(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 May 2006 (26.05.2006)

PCT

(10) International Publication Number
WO 2006/055040 A2

(51) International Patent Classification:
G01N 33/68 (2006.01) *C12N 15/90* (2006.01)
C12N 15/09 (2006.01)

(21) International Application Number:
PCT/US2005/019477

(22) International Filing Date: 3 June 2005 (03.06.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/628,948 19 November 2004 (19.11.2004) US

(71) Applicant *(for all designated States except US)*: **GOVERNMENT OF THE UNITED STATES OF AMERICA, DEPARTMENT OF HEALTH AND HUMAN SERVICES** [US/US]; 6011 Executive Boulevard, Suite 325, Rockville, MD 20852-3804 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **HARTLEY, James L.** [US/US]; 7409 Hillside Drive, Frederick, Maryland 21702 (US). **ESPOSITO, Dominic** [US/US]; 9305 White Rock Avenue, Frederick, Maryland 21702 (US). **STANARD, Kelly Jeanne** [US/US]; 9305 White Rock Avenue, Frederick, Maryland 21702 (US).

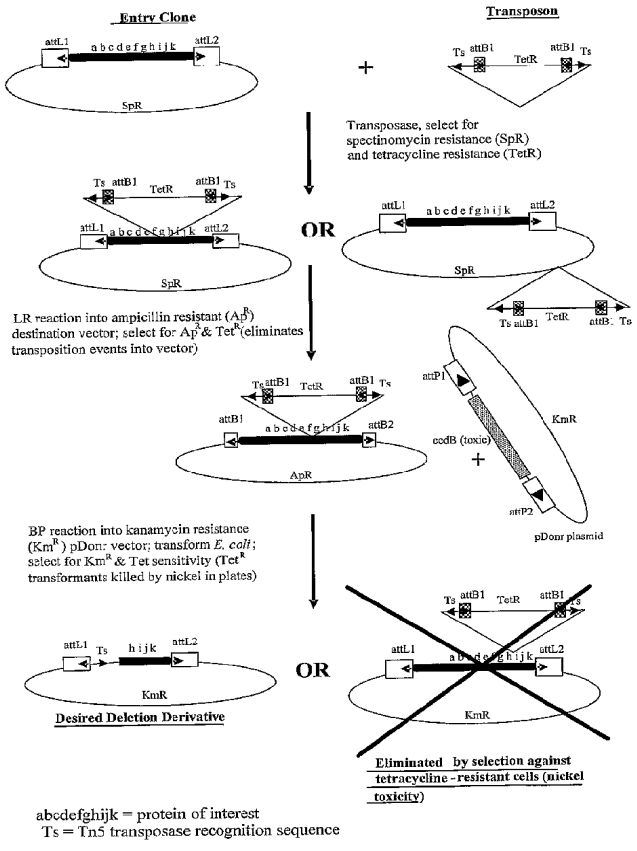
(74) Agent: **JAY, Jeremy**; 700 Thirteenth Street, N. W., Suite 300, Washington, DC 20005 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: IDENTIFICATION OF PROTEINS IN A GENOME



(57) Abstract: The invention provides a method for producing soluble proteins. In one embodiment, the invention provides a method of producing soluble deletion derivatives of a protein comprising transposon integration and recombinational cloning using site-specific recombinases. In another embodiment, the invention provides a method for identifying each of two or more soluble proteins, which comprises the simultaneous analysis of pooled open reading frames. Both embodiments comprise protein purification, protein separation, and mass spectroscopic analysis of the separated proteins. The invention also provides a transposon comprising a selectable marker gene flanked by identical inverted site-specific recombination sequences.

WO 2006/055040 A2

European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IIU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

IDENTIFICATION OF PROTEINS IN A GENOME

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This patent application claims the benefit of U.S. Provisional Patent Application No. 60/628,948, filed November 19, 2004.

FIELD OF THE INVENTION

[0002] This invention pertains to methods for identifying proteins in a sample.

BACKGROUND OF THE INVENTION

[0003] Drug discovery, a process by which bioactive compounds are identified and preliminarily characterized, is a critical step in the development of treatments for human diseases. Successful drug discovery depends upon the elucidation of the molecule or molecules in a human cell targeted by a particular candidate drug. When such targets are cellular proteins, the development of effective drugs requires that the structure of the protein targets be known. The primary bottleneck in protein structure identification is the difficulty in obtaining sufficient quantities of soluble proteins from mammalian cells. Indeed, when expressed recombinantly in heterologous hosts, many human proteins are expressed at low levels, are insoluble, or are soluble but difficult to purify. Obtaining sufficient quantities of soluble proteins for structural analysis, however, is difficult. Another difficulty is the lack of suitable methods for screening hundreds, or even thousands, of proteins for those that are expressed, soluble, and can be purified.

[0004] Thus, there remains a need for improved methods for identifying soluble proteins. Such methods will enable researchers to identify both individual protein targets of drugs, as well as protein families or protein signaling pathways, thereby enhancing drug development. The invention provides such methods. These and other advantages of the invention, as well as additional inventive features, will be apparent from the description of the invention provided herein.

BRIEF SUMMARY OF THE INVENTION

[0005] In one embodiment of the invention, a method for producing soluble deletion derivatives of a protein is provided. The method comprises (a) preparing a vector comprising a nucleic acid sequence encoding the protein, wherein the nucleic acid sequence is flanked by a first and a second site-specific recombination site, and wherein the first and second site-specific recombination sites do not recombine with each other, (b) incubating the vector of (a) in the presence of one or more transposons and a transposase protein under conditions sufficient to cause insertion of the one or more transposons into the vector,

wherein each of the one or more transposons comprises a third and a fourth site-specific recombination site, (c) transferring the nucleic acid sequence into a second vector, wherein the second vector comprises a fifth and a sixth site-specific recombination site, and propagating the second vector in a bacterium, (d) isolating one or more second vectors of (c) comprising the one or more transposons inserted into the nucleic acid sequence, (e) combining the second vector of (d) with a third vector to produce a mixture, wherein the third vector comprises a seventh and an eighth site-specific recombination site, and (f) incubating the mixture of (e) in the presence of at least one recombinase protein under conditions sufficient to cause recombination of (i) the seventh and eighth site-specific recombination sites of the third vector with (ii) the fourth site-specific recombination site of one transposon and the sixth site-specific recombination site of the second vector, wherein one or more nucleotides of the nucleic acid sequence are deleted.

[0006] In another embodiment of the invention, a method for identifying each of two or more soluble proteins is provided. The method comprises (a) preparing a mixture of two or more vectors each comprising a nucleic acid sequence encoding a soluble protein operatively linked to a promoter, (b) transferring each of the two or more vectors into one or more cells, (c) expressing the nucleic acid sequence in the one or more cells, wherein the two or more soluble proteins are produced, (d) purifying the two or more soluble proteins from the one or more cells, (e) separating the two or more soluble proteins, (f) isolating each of the two or more soluble proteins, and (g) subjecting the two or more soluble proteins to mass spectrometry, whereupon (i) each of the two or more soluble proteins is identified, and (ii) the amount of each of the two or more soluble proteins produced in the one or more cells is determined.

[0007] In yet a further embodiment, the above-described method of producing soluble deletion derivatives of a protein and the method of identifying two or more soluble proteins can be used alone or in combination.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] Figure 1 is a diagram illustrating a method of producing an amino terminal deletion derivative of a protein in accordance with the inventive method.

[0009] Figure 2 is a diagram illustrating a method of producing a carboxy terminal deletion derivative of a protein in accordance with the inventive method.

[0010] Figure 3 is a diagram illustrating a method of producing an amino terminal deletion derivative of a protein using a transposon comprising one site-specific recombination site in accordance with the inventive method.

[0011] Figure 4 is a diagram illustrating a method of producing a carboxy terminal deletion derivative of a protein using a transposon comprising one site-specific recombination site in accordance with the inventive method.

[0012] Figure 5 is a diagram illustrating a method of producing a double deletion derivative of a protein in accordance with the inventive method.

[0013] Figure 6 is an image of a 2D gel on which 688 affinity-purified *Caenorhabditis elegans* proteins were separated.

[0014] Figure 7 is an image of an SDS-PAGE gel demonstrating expression of each of twelve *C. elegans* ORFs identified by the inventive method. Each lane of the gel represents the total protein in each cell culture.

[0015] Figure 8 is an image of an SDS-PAGE gel demonstrating the soluble (S) and insoluble (I) proteins present in cells transformed with vectors containing each of twelve *C. elegans* ORFs identified by the inventive method.

[0016] Figure 9 is an image of an SDS-PAGE gel demonstrating expression of *C. elegans* ORF #8 induced with 1 mM IPTG (I), and in the uninduced (U) state.

[0017] Figure 10 is an image of an SDS-PAGE gel demonstrating the soluble (S) and insoluble (I) proteins produced by *C. elegans* ORF #8 as compared to a negative control (Neg).

[0018] Figure 11a is an image of an SDS-PAGE gel demonstrating protein expression in whole cells transduced with *C. elegans* ORFs identified in accordance with the inventive method.

[0019] Figure 11b is an image of an SDS-PAGE gel demonstrating protein expression in the soluble fraction of proteins expressed in cells transduced with *C. elegans* ORFs identified in accordance with the inventive method.

[0020] Figure 11c is an image of an SDS-PAGE gel demonstrating protein expression in the insoluble fraction of proteins expressed in cells transduced with *C. elegans* ORFs identified in accordance with the inventive method.

[0021] Figure 11d is an image of an SDS-PAGE gel demonstrating IMAC-purified soluble proteins expressed in cells transduced with *C. elegans* ORFs identified in accordance with the inventive method.

DETAILED DESCRIPTION OF THE INVENTION

[0022] The invention provides a method for producing soluble deletion derivatives of a protein. The method comprises (a) preparing a vector comprising a nucleic acid sequence encoding the protein, wherein the nucleic acid sequence is flanked by a first and a second site-specific recombination site, and wherein the first and second site-specific recombination sites do not recombine with each other, (b) incubating the vector of (a) in the

presence of one or more transposons and a transposase protein under conditions sufficient to cause insertion of the one or more transposons into the vector, wherein each of the one or more transposons comprises a third and a fourth site-specific recombination site, (c) transferring the nucleic acid sequence into a second vector, wherein the second vector comprises a fifth and a sixth site-specific recombination site, and propagating the second vector in a bacterium, (d) isolating one or more second vectors of (c) comprising the one or more transposons inserted into the nucleic acid sequence, (e) combining the second vector of (d) with a third vector to produce a mixture, wherein the third vector comprises a seventh and an eighth site-specific recombination site, and (f) incubating the mixture of (e) in the presence of at least one recombinase protein under conditions sufficient to cause recombination of (i) the seventh and eighth site-specific recombination sites of the third vector with (ii) the fourth site-specific recombination site of one transposon and the sixth site-specific recombination site of the second vector, wherein one or more nucleotides of the nucleic acid sequence are deleted.

[0023] The inventive method desirably is a high-throughput method in that it enables large-scale production of deletion derivatives (e.g., 100 or more, 500 or more, 1000 or more, or even 1,000 or more deletion derivatives). Typically and preferably, all or part of the method is automated.

[0024] The term "deletion derivative," as used herein, refers to a variant form of a full-length protein that contains a deletion of one or more amino acids from the amino acid sequence of the full-length protein. In the context of the invention, the one or more amino acids are deleted from the amino acid sequence of the full-length protein by deletion of one or more nucleotides from the nucleic acid sequence encoding the full-length protein. Any number of amino acids can be deleted from the full-length protein to produce the deletion derivative. By "soluble" is meant that the deletion derivative is capable of dissolving in a fluid, most preferably an aqueous buffer. The solubility of the deletion derivative can be determined using any suitable method known in the art, such as, for example, centrifugation- or filtration-based protein purification methods.

[0025] The inventive method comprises preparing vectors comprising a nucleic acid sequence encoding a protein. The term "vector" refers to a nucleic acid molecule that can transfer a heterologous nucleic acid insert contained therein from one organism to another. The vector desirably is a nucleic acid molecule (e.g., DNA or RNA). Preferably, the vector is comprised of DNA. The vector can be any suitable vector, such as a plasmid, a virus, a phage, an autonomously replicating sequence (ARS), or any other sequence that is capable of replicating *in vitro* or in a host cell. Typically and preferably, the vector is a plasmid that contains one or more restriction endonuclease sites which facilitate incorporation of the heterologous nucleic acid insert and heterologous regulatory sequences. Vectors can be

prepared using standard recombinant DNA and molecular biology techniques, such as those described in Sambrook et al., *Molecular Cloning, a Laboratory Manual*, 3rd edition, Cold Spring Harbor Press, Cold Spring Harbor, N.Y. (2001), Ausubel et al., *Current Protocols in Molecular Biology*, Greene Publishing Associates and John Wiley & Sons, New York, N.Y. (1994), Hartley et al., *Genome Res.*, *10*, 1788-95 (2000) and U.S. Patents 5,888,732, 6,270,969, 6,277,608, and 6,720,140. The inventive method comprises preparing a first vector comprising a nucleic acid sequence encoding the protein, wherein the nucleic acid sequence is flanked by a first and a second site-specific recombination site, and wherein the first and second site-specific recombination sites do not recombine with each other.

[0026] The nucleic acid sequence incorporated into the vector can encode any suitable protein from any suitable organism. In one embodiment, the protein is an animal protein. The protein can be derived from any suitable animal. Suitable animals include, for example, protozoa, echinoderms (e.g., sea urchin), annelids (e.g., earthworms), nematodes (e.g., *C. elegans*), mollusks, arthropods (e.g., crustaceans), insects, birds, amphibians, reptiles, and mammals (e.g., primates and rodents). Preferably, the organism is a mammal, and most preferably the organism is a human. In one embodiment, the nucleic acid sequence encodes a protein that is a target for a therapeutic agent (i.e., a drug). In this regard, the nucleic acid sequence can encode a protein that is a target for a known therapeutic agent, such as those described in, for example, *Physician's Desk Reference*, Medical Economics Co., Inc., Montvale, New Jersey (2004). Alternatively, the nucleic acid sequence can a protein that is not a target for a known drug, but which serves as a target against which a new drug is developed.

[0027] In a preferred embodiment of the invention, the nucleic acid sequence in the vector is flanked by a first and a second site-specific recombination site. The term "site-specific recombination site" refers to a specific sequence on a nucleic acid molecule that is recognized and bound by a specific recombination protein (also known as a recombinase), which facilitates the exchange (i.e., recombination) of nucleic acid sequences between two nucleic acid molecules. Numerous site-specific recombination systems from various organisms are known in the art, and include, for example, the integrase/*att* system derived from bacteriophage λ (see, e.g., Landy, *Curr. Opin. Genet. Devel.*, *3*, 699-707 (1993)), the Cre/*loxP* system derived from bacteriophage P1 (see, e.g., Hoess et al., In: *Nucleic Acids and Molecular Biology, vol. 4*, Eckstein and Lilley, eds., Springer-Verlag, Berlin, pp. 90-109 (1990)), and the FLP/FRT system derived from the *Saccharomyces cerevisiae* 2 μ circle plasmid (see, e.g., Broach et al., *Cell*, *29*, 227-234 (1982)). The first and second site-specific recombination sites can be any suitable site-specific recombination site, such as, for example a wild type *loxP* site and a mutant *loxP* site (see, e.g., Sauer, *Curr. Opin. Biotech.*,

5, 521-527 (1994)). Whatever site-specific recombinase is chosen, the first and second site-specific recombination sites do not recombine with each other.

[0028] In a preferred embodiment of the invention, the first and second site-specific recombination sites are derived from bacteriophage lambda, which uses site-specific recombination for integration into the *Escherichia coli* chromosome (see, e.g., Landy et al., *Ann. Rev. Biochem.*, *58*, 913-949 (1989). Site-specific recombination occurs between site-specific attachment (*att*) sites. Specifically, site-specific recombination occurs between an *attB* site on the *E. coli* chromosome and an *attP* site on the lambda chromosome. Upon lambda integration, recombination occurs between *attB* and *attP* sites to give rise to *attL* and *attR* sites. Preferably, the first and second site-specific recombination sites flanking the nucleic acid sequence are each an *attL* site.

[0029] The inventive method further comprises incubating a first vector in the presence of one or more transposons and a transposase protein under conditions sufficient to cause insertion of the transposon into the vector. The term "transposon," as used herein, refers to a mobile genetic element. Transposons are structurally variable, and can encode a transposition catalyzing enzyme, called a transposase, flanked by DNA sequences organized in inverted orientations (see, e.g., Sherratt, ed., *Mobile Genetic Elements*, Oxford University Press (1995), and Berg and Howe, eds., *Mobile DNA*, American Society for Microbiology, Washington, DC (1989)). Transposons are used in the art to insert DNA into target DNA sequences. Transposons typically contain additional genes not related to transposition, such as antibiotic resistance genes, and are classified as "simple" (class 1) and "complex" (class 2). As a general rule, the insertion of transposons into target DNA is a random event, but transposon insertion can also occur with some sequence specificity. For example, transposon Tn7 can integrate itself into a specific site in the *E. coli* genome as one part of its life cycle (see, e.g., Stellwagen et al., *Trends in Biochemical Sciences*, *23*, 486-490 (1998)). Moreover, the Tn5 transposon is known to preferentially insert at GC rich sequences (see, e.g., Herron et al., *Nucl. Acids Res.*, *32*, e113, (2004)). Any suitable transposon can be used in the inventive method. Suitable transposons include, for example, Tn7, TnS, Tn1, Tn3, Tn4, Tn7, Tn501, Tn551, Tn9, Tn903, Tn1681, Tn10, and Tn5. One of ordinary skill in the art will appreciate that incubation of the first vector with the one or more transposons can occur *in vitro* or *in vivo* (e.g., in a bacterium). In a preferred embodiment of the invention, the first vector and the one or more transposons are incubated *in vitro*.

[0030] Each of the one or more transposons preferably is a Tn5 transposon. Because transposition of Tn5 typically shows some sequence selectivity in target site preference, the use of multiple transposons that vary in reading frame can be employed in the inventive method. In this regard, any number of transposons can be employed in the inventive method. Preferably, the first vector is incubated with at least one transposon (e.g., 1, 3, 5, or

more transposons). Most preferably, the first vector is incubated with three transposons, each of which comprises a different reading frame (e.g., two additional transposons having one or two extra nucleotides at the 3' ends). The use of three transposons can yield a viable in-frame deletion in accordance with the inventive method. Whereas a single transposon may insert frequently at a location out of frame, three transposons inserting at the same site will improve the likelihood of at least one in-frame deletion.

[0031] In accordance with the inventive method, the transposon comprises a third and a fourth site-specific recombination site, each of which can be any suitable site-specific recombination site such as those described herein. Preferably, the third and the fourth site-specific recombination sites are derived from phage lambda. More preferably, the third and fourth site-specific recombination sites are each an *attB* site. The third and fourth site-specific recombination sites can be different *attB* sites (e.g., an *attB3* site and an *attB4* site) (see, e.g., Cheo et al., *Genome Res.*, *14*, 2111-20 (2004)), but preferably they are identical and inverted. In addition to the third and fourth site-specific recombination sites, the transposon can comprise other elements that facilitate identification of cells that contain the transposon. Such elements are typically referred to in the art as "selectable marker genes," and include, for example, antibiotic resistance genes, genes encoding protein products which are otherwise lacking in a recipient cell (e.g., herpes simples virus thymidine kinase (HSV-TK)), and genes encoding protein products that can be visually identified (e.g., β-galactosidase, luciferase, and green fluorescent protein (GFP)). The transposon can comprise any suitable number of selectable marker genes. Preferably, the transposon comprises one selectable marker gene.

[0032] In a preferred embodiment of the invention, the transposon comprises an antibiotic resistance gene. The transposon can comprise any suitable antibiotic resistance gene. Suitable antibiotic resistance genes are known in the art and include, for example, ampicillin resistance, kanamycin resistance, chloramphenicol resistance, tetracycline resistance, spectinomycin resistance, streptomycin resistance, and sulfonamide resistance genes. In addition to antibiotic resistance genes, the transposon can further comprise a nucleic acid sequence that encodes a protein product that is toxic to cells harboring the transposon when grown under specific conditions. In this manner, cells containing the transposon are killed (i.e., "selected against") when grown under such conditions. For example, expression of a tetracycline resistance gene also confers sensitivity to nickel (see, e.g., Podolsky et al., *Plasmid*, 36, 112-115 (1996)). Thus, when grown in medium containing nickel (e.g. nickel sulphate), cells containing a tetracycline resistance gene do not survive, and are selected against. In addition, mutant phenylalanine tRNA synthetase (PheS) protein causes the incorporation of chlorophenylalanine (Cl-Phe) into proteins, which is toxic to cells. Thus, when grown in medium containing Cl-Phe, cells containing a

mutant PheS gene do not survive, and are thus selected against (see, e.g., Kast, *Gene*, *28*, 109-114 (1994)). In a preferred embodiment of the invention, the transposon comprises a kanamycin resistance gene (kan^R) or a tetracycline resistance gene (tet^R) (see, e.g., Podolsky et al., *supra*).

[0033] The first vector can be incubated with the transposon and the transposase using any suitable molecular biology technique known in the art, so long as the conditions under which incubation occurs are sufficient to cause insertion of the transposon into the vector. Suitable methods are disclosed in, for example, Miller et al., eds., *Mobile Genetic Elements: Protocols and Genomic Applications (Methods in Molecular Biology)*, Humana Press, Totowa, New Jersey (2004). As a result of incubation of the first vector with the transposon and transposase, the transposon integrates into the vector. In accordance with the inventive method, two types of integration events are generated. First, the transposon can integrate within the nucleic acid sequence in the vector that encodes the protein of interest. Alternatively, the transposon can integrate within the vector but not within the nucleic acid sequence encoding the protein of interest.

[0034] The invention further comprises transferring the nucleic acid sequence into a second vector, and propagating the second vector in a bacterium, wherein the second vector comprises a fifth and a sixth site-specific recombination site. The second vector carries a selectable marker gene, and preferably carries a different selectable marker gene (e.g., a different antibiotic resistance gene) than the selectable marker gene(s) carried by the transposon and by the first vector. Bacterial cells are then transfected with the second vector, and grown under conditions such that only those cells containing both the second vector, and the transposon integrated therein, are able to grow. In a preferred embodiment, bacterial cells are grown in the presence of an antibiotic against which the transposon is resistant, and an antibiotic against which the second vector is resistant. Any suitable bacterium known in the art may be used to propagate the second vector. Preferably, the bacterium used to propagate the second vector is *E. coli*. The second vector can be introduced into the bacterium and propagated therein using routine molecular biology and recombinant DNA techniques, such as, for example, chemical transformation or electroporation (see, e.g., Sambrook et al., *supra*).

[0035] The nucleic acid sequence, and transposon integrated therein, can be transferred to the second vector using routine recombinant DNA and molecular biology techniques, such as those known in the art and described herein. Preferably, the nucleic acid sequence is transferred to the second vector utilizing site-specific recombination provided by the Gateway® Technology cloning system (Invitrogen Life Technologies, Carlsbad, CA). Specifically, in a preferred embodiment, when the first and second site-specific recombination sites flanking the nucleic acid sequence are each an *attL* site, the first vector

is combined with a second vector containing two *attR* site-specific recombination sites, as well as an ampicillin resistance gene, in the presence of a suitable recombinase protein. In this manner, recombination between the *attL* and *attR* sites results in transfer of the nucleic acid sequence to the second vector, and the creation of *attB* site-specific recombination sites flanking the nucleic acid sequence. The Gateway® Technology is further described in, for example, the instruction manual entitled "*Gateway Technology (Version E),*" available from Invitrogen Corp., Life Technologies Division, Carlsbad, CA, and in Hartley et al., *supra,* and U.S. Patents 5,888,732, 6,270,969, 6,277,608, and 6,720,140.

[0036] Following transfer of the nucleic acid sequence, and transposon integrated therein, to the second vector, the inventive method further comprises isolating the second vector comprising the transposon inserted into the nucleic acid sequence. The second vector can be isolated using routine molecular biology techniques, such as those described in Ausubel et al., *supra,* Sambrook et al., *supra.*

[0037] The inventive method further comprises combining the second vector with a third vector to produce a mixture. In a preferred embodiment of the invention, the third vector comprises a seventh and an eighth site-specific recombination site. The seventh and eighth site-specific recombination sites each can be any suitable site-specific recombination sites described herein. As discussed above, recombination between the *attL* sites on the first vector and the *attR* sites on the second vector creates *attB* sites on the second vector. Thus, to facilitate recombination between the second vector and the third vector, the seventh and eighth site-specific recombination sites on the third vector are preferably each an *attP* site-specific recombination site. To allow for selection and isolation of the desired deletion derivative, the third vector preferably carries a selectable marker gene (e.g., an antibiotic resistance gene). The third vector can comprise any suitable selectable marker gene, so long as the selectable marker gene of the third vector is different than the selectable marker genes carried by the transposon and the second vector. In one embodiment, for example, the third vector comprises a kanamycin resistance gene as a selectable marker gene.

[0038] As discussed above, the seventh and eighth site-specific recombination sites of the third vector preferably are *attP* site-specific recombination sites (e.g., *attP1* and *attP2*), the fifth and sixth site-specific recombination sites of the second vector, which flank the nucleic acid sequence, preferably are *attB* site-specific recombination sites (e.g., *attB1* and *attB2*), and the third and fourth site-specific recombination sites on the transposon preferably are *attB* site-specific recombination sites (e.g., two inverted *attB1* sites). Thus, the mixture of the second and third vectors are incubated in the presence of at least one recombinase protein under conditions sufficient to cause recombination of the *attP* sites on the third vector with one *attB* site on the second vector, and one *attB* site on the transposon.

[0039] For example, deletion of the amino terminus of a protein of interest can be performed as illustrated in Figure 1. In this embodiment, the third, fourth, and fifth site-specific recombination sites are each an *attB1* site, and the sixth site-specific recombination site is an *attB2* site. The seventh site-specific recombination site on the third vector is an *attP1* site, while the eighth site-specific recombination site on the third vector is an *attP2* site. In this manner, an *attB1* site will recombine with an *attP1* site of appropriate orientation, and an *attB2* site will recombine with an *attP2* site of appropriate orientation. Thus, two classes of recombination events are generated: the seventh and eighth site-specific recombination sites of the third vector can recombine with (a) the fourth site-specific recombination site of the transposon and the sixth site-specific recombination site of the second vector, respectively, or (b) the fifth and sixth site-specific recombination sites of the second vector, respectively. Recombination of the seventh and eighth site-specific recombination sites of the third vector with the fourth site-specific recombination site of the transposon and the sixth site-specific recombination site of the second vector, respectively, is preferred when generating an amino terminal deletion derivative of a protein of interest. Such a recombination event transfers the portion of the nucleic acid sequence encoding the carboxy portion of the protein into a separate vector, eliminating the transposon but retaining one transposase recognition sequence. The method set forth in Figure 1, however, is merely an exemplary embodiment of the present invention, and should not be construed as limiting the scope of the inventive method.

[0040] Deletion of the carboxy terminus of a protein of interest can be performed as illustrated in Figure 2. Specifically, the third, fourth, and sixth site-specific recombination sites can each be an *attB2* site, while the fifth site-specific recombination site can be an *attB1* site. The seventh and eighth site-specific recombination sites of the third vector can be an *attP1* site and an *attP2* site, respectively. Thus, as discussed above, an *attB1* site will recombine with an *attP1* site of appropriate orientation, and an *attB2* site will recombine with an *attP2* site of appropriate orientation. In this manner, two classes of recombination events are generated: the seventh and eighth site-specific recombination sites of the third vector can recombine with (a) the fifth site-specific recombination site of the second vector and the third site-specific recombination site of the transposon, respectively, or (b) the fifth and sixth site-specific recombination sites of the second vector, respectively.
Recombination of the seventh and eighth site-specific recombination sites of the third vector with the fifth site-specific recombination site of the second vector and the third site-specific recombination site of the transposon, respectively, is preferred when generating a carboxy terminal deletion derivative of a protein of interest. Such a recombination event transfers the portion of the nucleic acid sequence encoding the amino portion (i.e., fragment) of the protein into a separate vector, eliminating the transposon. The method set forth in Figure 2,

however, is merely an exemplary embodiment of the present invention, and should not be construed as limiting the scope of the inventive method.

[0041] While the transposon preferably comprises two site-specific recombination sites (i.e., a third and fourth site-specific recombination site), a transposon comprising only one site-specific recombination site (i.e., a third site-specific recombination site) also is within the scope of the inventive method. In such embodiments, the second vector comprises a fourth and a fifth site-specific recombination site, while the third vector comprises a sixth and a seventh site-specific recombination site. The sixth and seventh site-specific recombination sites of the third vector preferably are *attP* site-specific recombination sites (e.g., *attP1* and *attP2*), the fourth and fifth site-specific recombination sites of the second vector, which flank the nucleic acid sequence, preferably are *attB* site-specific recombination sites (e.g., *attB1* and *attB2*), and the third site-specific recombination site on the transposon preferably is an *attB* site-specific recombination site (e.g., *attB1* or *attB2*). Depending upon the identity and orientation of the third site-specific recombination site on the transposon, a deletion of nucleic acids encoding the amino or the carboxy end of the protein of interest can be generated, as illustrated in Figures 3 and 4.

[0042] As discussed above, the third vector and the transposon each possesses one or more selectable marker gene(s) that allow selection of bacterial cells containing a third vector that harbors a desired deletion derivative. In one embodiment, the transposon carries a tetracycline resistance gene, while the third vector carries a kanamycin resistance gene. Thus, the desired recombination event is selected for by introducing the recombination products into suitable bacterial cells, and growing such cells in the presence of nickel and kanamycin. In this manner, cells that are resistant to kanamycin and nickel, i.e., cells which have lost the transposon comprising the tetracycline resistance gene and part of the nucleic acid sequence as a result of recombination, are selected for.

[0043] In accordance with the invention, recombination of the seventh and eighth site-specific recombination sites of the third vector with the fourth site-specific recombination site of the transposon and the sixth site-specific recombination site of the second vector, respectively, results in the transfer of a portion of the nucleic acid sequence into the third vector, which is equivalent to deletion of one or more nucleotides of the nucleic acid sequence. When grown in a suitable host (e.g., a bacteria), the nucleic acid sequence is expressed, and a soluble deletion derivative of the protein of interest is produced. By virtue of the integration of the one or more transposons independently into two or more vectors, each comprising the same nucleic acid sequence, different deletion derivatives of the same nucleic acid sequence are produced. Alternatively, by virtue of the integration of the one or more transposons into two or more vectors, each of the vectors comprising two or more different nucleic acids sequences, deletion derivatives of the two or more nucleic acid

sequences are produced. Moreover, depending upon the identity and orientation of the third and fourth site-specific recombination sites in the transposon, the inventive method can result in deletion of amino acids at the amino terminus (N-terminus) or at the carboxy terminus (C-terminus) (see Figures 1 and 2). The inventive method of producing two or more soluble deletion derivatives of a protein can be used to generate deletions at both the N-terminus and the C-terminus ("double deletions"). In this regard, a nucleic acid sequence encoding a protein of interest is subjected to two successive cycles of the inventive method of producing one or more soluble deletion derivatives. For example, an N-terminal deletion derivative of a protein of interest generated in accordance with the inventive method can be subjected to a second cycle of the inventive method to generate a C-terminal deletion of the protein, and vice versa (see Figure 5).

[0044] Once one or more desired deletion derivatives are generated, the deletion derivative(s) can be sequenced, screened for solubility, and/or screened for proper reading frame. In this regard, the nucleic acid sequence encoding the deletion derivative can be sequenced using DNA sequencing methods known in the art. By sequencing the deletion derivative, the exact deletion endpoint can be determined by comparing the nucleic acid sequence of the deletion derivative to the nucleic acid sequence encoding the protein from which the deletion derivative is derived. The solubility of the deletion derivative(s) also can be determined using methods known in the art, such as gel electrophoresis or mass spectrometry. In addition, the solubility of the deletion derivative(s) can be determined by screening specially-designed expression vectors encoding carboxy-terminal fusion proteins of the deletion derivative(s) of interest and a fluorescent protein or an antibiotic resistance gene.

[0045] One of ordinary skill in the art will appreciate that the nucleic acid sequences encoding deletion derivatives generated by the inventive method must be in the correct reading frame, so as to produce a soluble deletion derivative protein. Thus, the nucleic acid sequences encoding the deletion derivatives generated in accordance with the inventive method preferably are screened for the proper reading frame. Suitable screening methods are known in the art, and include, for example, the use of software packages that translate a given nucleic acid sequence into all possible open reading frames (e.g., Transeq, European Bioinformatics Institute) and compare open reading frame sequences to sequences of known proteins (e.g., BLAST). In addition, a nucleic acid sequence encoding a deletion derivative can be engineered to comprise a selectable marker gene as part of a fusion protein (e.g., an antibiotic resistance gene), such that selection of the phenotype of the selectable marker gene will co-select for deletion derivatives in the proper reading frame.

[0046] The invention further provides a method for identifying one or more protein portions (i.e., fragments) in a sample. The method comprises (a) preparing a mixture of two

or more (e.g., 2, 10, 50, 100, 500, or 1000 or more) soluble deletion derivatives of one or more proteins in accordance with the inventive method, (b) separating the two or more soluble deletion derivatives, (c) isolating each of the two or more separated soluble deletion derivatives, and (d) subjecting each of the two or more soluble deletion derivatives to mass spectrometry, whereupon the one or more protein portions is identified.

[0047] Separation of the two or more soluble deletion derivatives can be performed using any suitable method known in the art. The two or more soluble deletion derivatives preferably are produced in and isolated from host cells (e.g., bacterial cells) prior to separation. In this respect, cells preferably are harvested and lysed using routine molecular biology techniques. Insoluble proteins are then removed from the cell lysate using any suitable method, such as, for example, centrifugation, and can be further processed in accordance with other embodiments of the invention. The soluble deletion derivatives are then purified from the cell lysate, by, for example, affinity purification. Methods for isolating and purifying recombinantly produced proteins are described in, for example, Sambrook et al., *supra*, and Ausubel et al., *supra*. Examples of suitable protein separation methods include, but are not limited to, centrifugation, ion exchange chromatography, isoelectric focusing, reversed-phase liquid chromatography, and gel electrophoresis. Preferably, the two or more soluble deletion derivatives are separated using gel electrophoresis (e.g., one-dimensional or two-dimensional gel electrophoresis). Most preferably, the two or more soluble deletion derivatives are separated using two-dimensional gel electrophoresis (2DGE). 2DGE typically involves separation of proteins in a first dimension by charge using isoelectric focusing (IEF). The charge-focused proteins are then separated in a second dimension according to size by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) (see, e.g., Lin et al., *Biochimica et Biophysica Acta*, *1646*, 1-10 (2003) and Ong et al., *Biomol. Eng.*, *18*, 195-205 (2001)). The separated proteins can be digested with a protease (e.g., trypsin) to aid in protein sequencing. In a preferred embodiment, protein digestion is performed in the gel once the two or more soluble deletion derivatives have been separated. Following separation, the two or more proteins can be visualized by staining the gel using any suitable staining reagent (e.g., Coomassie Blue), and analyzed to compare the relative intensities of each protein "spot" on the gel.

[0048] Following separation of the two or more soluble deletion derivatives, each of the two or more deletion derivatives is isolated from the separation medium. The deletion derivatives can be isolated using any suitable technique, such as by extracting the protein "spots" from the gel. Extraction of protein spots from a gel typically involves the physical cutting of the spot from the gel.

[0049] To identify the two or more soluble deletion derivatives, the isolated deletion derivatives preferably are subjected to mass spectrometry. In mass spectrometry, a

substance is ionized, and the positive fragments which are produced (cations and radical cations) are accelerated in a vacuum through a magnetic field and are sorted on the basis of mass-to-charge ratio (m/z). Since the bulk of the ions produced in the mass spectrometer carry a unit positive charge, the value m/z is equivalent to the molecular weight of the fragment. Any suitable mass spectrometry method can be used in connection with the inventive method. Examples of suitable mass spectrometry methods include matrix-assisted laser desorption/ionization mass spectrometry (MALDI), matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) mass spectrometry, plasma desorption/ionization mass spectrometry (PDI), electrospray ionization mass spectrometry (ESI), surface enhanced laser desorption/ionization-time of flight (SELDI-TOF) mass spectrometry, and liquid chromatography mass spectrometry (LC-MS or LC-MS-MS). In time-of-flight (TOF) methods of mass spectrometry, charged (ionized) molecules are produced in a vacuum and accelerated by an electric field produced by an ion-optic assembly into a free-flight tube or drift time. The velocity to which the molecules may be accelerated is proportional to the square root of the accelerating potential, the square root of the charge of the molecule, and inversely proportional to the square root of the mass of the molecule. The charged molecules travel down the TOF tube to a detector. In LC-MS-MS, liquid chromatography is used to separate the two or more deletion derivatives, which are then subject to mass spectrometry. Computer software selects the fragments from the mass spectrometer, channels them for ionization via collision-induced dissociation (CID), and a second mass spectrometer detects the fragments. The fragmentation caused by these collisions gives fragments with a "ladder" of masses, each "rung" of which corresponds to the next amino acid in the chain. As such, the difference in masses between the fragments provides the identity of each amino acid, and thus the sequence of the deletion derivative. Mass spectrometry methods are further described in, for example, International Patent Application Publication No. WO 93/24834, U.S. Patent 5,792,664, U.S. Patent Application Publication No. 2004/0033530 A1, and Hillenkamp et al., *Matrix Assisted UV-Laser Desorption/Ionization: A New Approach to Mass Spectrometry of Large Biomolecules, Biological Mass Spectrometry*, Burlingame and McCloskey, eds., Elsevier Science Publ., pp. 49-60 (1990).

[0050] The one or more nucleic acid sequences can be derived from any suitable organism, and need not be derived from the same organism. In this regard, the inventive method for identifying one or more protein portions in a sample can be used to identify one or more proteins from multiple different organisms. Alternatively, the inventive method for identifying one or more protein portions in a sample can be used to identify one or more proteins derived from the same organism. Typically and preferably, the organism is an animal. The two or more proteins can be derived from any suitable animal. Suitable

animals include, for example, protozoa, echinoderms (e.g., sea urchin), annelids (e.g., earthworms), nematodes (e.g., *C. elegans*), mollusks, arthropods (e.g., crustaceans), insects, birds, amphibians, reptiles, and mammals (e.g., primates and rodents). Preferably, the organism is a mammal, and most preferably the organism is a human.

[0051] The invention further provides a method for identifying two or more soluble proteins, or peptide fragments thereof. The method comprises (a) preparing a mixture of two or more vectors each comprising a nucleic acid sequence encoding a soluble protein operatively linked to a promoter, (b) transferring each of the two or more vectors into one or more cells, (c) expressing the nucleic acid sequence in each of the one or more cells, wherein the two or more soluble proteins are produced, (d) purifying the two or more soluble proteins from the one or more cells, (e) separating the two or more soluble proteins, (f) isolating each of the two or more soluble proteins, and (g) subjecting the two or more soluble proteins to mass spectrometry, whereupon the each of the two or more soluble proteins is identified, and the amount of each of the two or more soluble proteins produced in the one or more cells is determined. Descriptions of the vector, soluble protein, and mass spectrometric analysis set forth above in connection with other embodiments of the invention also are applicable to those same aspects of the aforesaid inventive method for identifying each of two or more soluble proteins.

[0052] The inventive method for each of two or more soluble proteins can be used to identify any number of proteins (e.g., 2, 10, 20, 50, 100, or 1000 or more) simultaneously. The method comprises preparing two or more vectors (e.g., 2, 10, 20, 50, 100, or 1000 or more), comprising a nucleic acid sequence encoding a soluble protein operatively linked to a promoter. In the context of the inventive method, each vector comprises a nucleic acid sequence encoding one soluble protein. The nucleic acid is operatively linked to (i.e., under the transcriptional control of) one or more promoter and/or enhancer elements. Techniques for operatively linking sequences together are well known in the art. A "promoter" is a DNA sequence that directs the binding of RNA polymerase and thereby promotes RNA synthesis. A promoter can be native or non-native to the nucleic acid sequence to which it is operably linked.

[0053] Any promoter (i.e., whether isolated from nature or produced by recombinant DNA or synthetic techniques) can be used in connection with the invention to provide for transcription of the nucleic acid sequence. The promoter preferably is capable of directing transcription in a cell (e.g., a prokaryotic or eukaryotic cell). The functioning of the promoter can be altered by the presence of one or more enhancers and/or silencers present on the vector. "Enhancers" are cis-acting elements of DNA that stimulate or inhibit transcription of adjacent genes. An enhancer that inhibits transcription also is termed a "silencer." Enhancers differ from DNA-binding sites for sequence-specific DNA binding

proteins found only in the promoter (which also are termed "promoter elements") in that enhancers can function in either orientation, and over distances of up to several kilobase pairs (kb), even from a position downstream of a transcribed region.

[0054] Any suitable promoter or enhancer sequence can be used in the context of the invention. In this respect, the nucleic acid sequence can be operatively linked to a constitutive promoter. Any suitable constitutive promoter can be used in connection with the inventive method. Suitable constitutive promoters include, for example, cytomegalovirus (CMV) promoters, such as the CMV immediate-early promoter (described in, for example, U.S. Patents 5,168,062 and 5,385,839), promoters derived from human immunodeficiency virus (HIV), Rous sarcoma virus (RSV) promoters, such as the RSV long terminal repeat, mouse mammary tumor virus (MMTV) promoters, HSV promoters, such as the Lap2 promoter or the herpes thymidine kinase promoter (Wagner et al., *Proc. Natl. Acad. Sci., 78,* 144-145 (1981)), promoters derived from SV40 or Epstein Barr virus, the YY1 promoter, the ubiquitin promoter, and the like.

[0055] Preferably, the promoter is a regulatable promoter, i.e., a promoter that is up- and/or down-regulated in response to appropriate signals. Suitable regulatable promoter systems include, but are not limited to, the tetracycline expression system, the IL-8 promoter, the metallothionine inducible promoter system, the bacterial lacZYA expression system, and the T7 polymerase system. Further, promoters that are selectively activated at different developmental stages (e.g., globin genes are differentially transcribed from globin-associated promoters in embryos and adults) can be employed. The promoter sequence can contain at least one regulatory sequence responsive to regulation by an exogenous agent. The regulatory sequences are preferably responsive to exogenous agents such as, but not limited to, drugs, hormones, radiation, sugars, salts, and the like.

[0056] The promoter can be a cell-specific or tissue-specific promoter, i.e., a promoter that is preferentially activated in a given cell or tissue and results in expression of a gene product in the tissue where activated. A cell-specific or tissue-specific promoter suitable for use in the invention can be chosen by the ordinarily skilled artisan based upon the target tissue or cell-type.

[0057] In a preferred embodiment of the invention, the two or more vectors are transferred into one or more cells. In a preferred embodiment, each cell is transduced with only one vector. The two or more vectors can be transferred into one or more cells using any suitable method known in the art. Suitable methods include, for example, chemical transformation, electroporation, high velocity bombardment with DNA-coated microprojectiles, incubation with calcium phosphate-DNA precipitate, direct microinjection into single cells, calcium phosphate or DEAE-dextran-mediated transfection, polybrene transfection, protoplast fusion, liposome-mediated transfection, and the like (see, e.g.,

Sambrook et al., *supra*). In an alternative embodiment, each of the two or more vectors need not be transferred to two or more cells in a one-to-one ratio. For example, one or more vectors can be transferred into a single cell.

[0058] The two or more vectors can be transferred into any cells capable of transduction which permit expression of the nucleic acid sequence encoding the soluble protein and can be readily propagated (i.e., cultured). Examples of suitable cells include bacterial cells, such as *E. coli*, yeast cells (e.g. *Pichia pastoris* and *Saccharomyces cerevisiae*), and animal cells, such as insect cells and *C. elegans* cells, and mammalian cloned cells, such as HeLa cells, CHO cells, and VERO cells. Preferably the one or more cells are each bacterial cells. The two or more vectors can be introduced into the one or more cells as described herein, and the one or more cells can be combined to form a single sample. In a preferred embodiment, the two or more vectors are combined into a single sample, and the mixture of vectors is transduced into host cells.

[0059] The inventive method further comprises expressing the nucleic acid sequence in each of the one or more cells, wherein the two or more soluble proteins are produced. As described herein, expression of a nucleic acid sequence can be regulated (e.g., induced) when the nucleic acid sequence is operatively linked to a regulatable promoter. Preferably, the nucleic acid sequence is operatively linked to an inducible promoter. The nucleic acid sequence can be operatively linked to any suitable inducible promoter, such as those described herein. In the presence of an appropriate signal, the inducible promoter is activated, and the nucleic acid sequence encoding the soluble protein is expressed.

[0060] Expression of each of the two or more nucleic acids results in production of the two or more soluble proteins, or peptide fragments thereof. In accordance with the inventive method, each of the two or more soluble proteins is purified from the one or more cells. The two or more soluble proteins can be purified using any suitable method known in the art, such as those described herein. Specifically, cells preferably are harvested and lysed using routine molecular biology techniques. Insoluble proteins are then removed from the cell lysate using any suitable method, such as, for example, centrifugation, and can be further processed in accordance with other embodiments of the invention. The soluble proteins are then purified from the cell lysate, by, for example, affinity purification. Methods for isolating and purifying recombinantly produced proteins are described in, for example, Sambrook et al., *supra*, and Ausubel et al., *supra.*

[0061] Once purified, the inventive method further comprises separating the two or more soluble proteins, isolating each of the two or more soluble proteins, and subjecting the two or more soluble proteins to mass spectrometry. The two or more soluble proteins can be separated using any suitable technique known in the art. Preferably, the two or more soluble proteins are separated by electrophoresis, more preferably two-dimensional gel

electrophoresis (2DGE). The separated proteins can be digested with a protease (e.g., trypsin) to aid in protein sequencing. In a preferred embodiment, protein digestion is performed in the gel once the two or more soluble deletion derivatives have been separated.

[0062] Following separation of the two or more soluble proteins, each of the two or more soluble proteins is isolated from the separation medium. The soluble proteins can be isolated using any suitable technique, such as by extracting the protein "spots" from the gel. Extraction of protein spots from a gel typically involves the physical cutting of the spot from the gel.

[0063] Once isolated and purified, each of the two or more soluble proteins, or peptide fragments thereof, is identified by subjecting the two or more soluble proteins to mass spectrometry. Any suitable mass spectrometry method can be used in connection with the inventive method. Suitable mass spectrometry methods are described herein and include, for example, matrix-assisted laser desorption/ionization mass spectrometry (MALDI), matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF or MALDI-TOF-TOF) mass spectrometry, plasma desorption/ionization mass spectrometry (PDI), electrospray ionization mass spectrometry (ESI), surface enhanced laser desorption/ionization-time of flight (SELDI-TOF) mass spectrometry, and liquid chromatography mass spectrometry (LC-MS or LC-MS-MS). In a preferred embodiment, the two or more soluble proteins, or peptide fragments thereof, are subjected to MALDI-TOF-TOF or LC-MS-MS mass spectrometry.

[0064] One of the goals of protein mass spectrometry is protein sequencing. Thus, the inventive method also encompasses the identification of the sequence of each of the two or more soluble proteins, or peptide fragments thereof. In this respect, the proteins can be sequenced using any suitable method known in the art. Suitable protein sequencing methods include, for example, Edman sequencing (see, e.g., Gevaert et al., *Electrophoresis*, *21*, 1145-1154 (2000)), and sequencing by mass spectrometry (see, e.g., Kinter and Sherman, *Protein Sequencing and Identification Using Tandem Mass Spectrometry*, Wiley-Interscience (2000), and U.S. Patents 6,632,339 and 6,706,529). Mass spectrometry-based protein identification methods include, for example, MALDI-MS peptide mass fingerprinting (MALDI-MS-PMF), MALDI-MS post-source decay analysis (MALDI-MS-PSD), and liquid chromatography mass spectrometry (LC-MS or LC-MS-MS). In MALDI-MS-PMF, a protein of interest, which typically is purified by 2-D gel electrophoresis, is either enzymatically or chemically cleaved and an aliquot of the obtained peptide mixture is analyzed by mass spectrometric techniques, thereby generating a mass "fingerprint" of the protein. The mass fingerprint is subsequently compared to "virtual" fingerprints obtained by theoretical cleavage of protein sequences stored in databases (e.g., MOWSE, ProFound, PeptIdent, and PeptideSearch), and the top scoring proteins are retrieved as possible

candidate proteins (see, e.g., Gevaert et al., *supra*). PSD fragments are generated in the field-free drift region after MALDI. It has been postulated that PSD primarily is the result of amide bond cleavages. Due to the high complexity of PSD spectra, however, MALDI-MS-PSD is not frequently used in the art to identify proteins (see, e.g., Gevaert et al., *supra*). As discussed above, LC-MS-MS involves mass spectroscopic analysis of proteins separated by liquid chromatography. Protein fragments selected using computer software are channeled for ionization via collision-induced dissociation (CID), and a second mass spectrometer detects the fragments. The fragmentation caused by these collisions gives fragments with a "ladder" of masses, each "rung" of which corresponds to the next amino acid in the chain. As such, the difference in masses between the fragments provides the identity of each amino acid, and thus the sequence of the protein. Upon preliminary identification of a protein in a sample, the identity of the protein can be confirmed by various protein detection methods known in the art, such as, for example, enzyme-linked immunosorbent assay (ELISA), Western blot analysis, immunoprecipitation, and isoelectric focusing followed by 2-D gel electrophoresis.

[0065] The amount of each of the two or more soluble proteins, or peptide fragments thereof, produced in the one or more cells can be determined prior to or following mass spectrometry of the two or more soluble proteins, or peptide fragments thereof. In one embodiment, the intensity of the "spots" on the 2D gel corresponding to each of the two or more soluble proteins analyzed by mass spectrometry can be measured. In this regard, the intensity of a particular protein spot on the 2D gel is approximately proportional to the amount of that protein present in a cell. The protein "spots" on the 2D gel can be analyzed using various commercially-available computer software packages, such as, for example, PDQuest 7.3.0 Software (Bio-Rad Laboratories, Hercules, CA), Phoretix 2D Gel Image Analysis Software (United Bioinformatica Inc., Calgary, AB, Canada), and ProFINDER 2D Image Analysis Software (PerkinElmer Life & Analytical Science, Inc., Boston, MA).

[0066] The method of producing soluble deletion derivatives of a protein and the method of identifying two or more soluble proteins described herein can be used alone or in combination to maximize the ability of one of ordinary skill in the art to identify proteins with potential therapeutic significance. Insoluble proteins that are not identified by the method of identifying two or more soluble proteins can be treated with solubilizing agents and used to generate soluble deletion derivatives as described herein. Such deletion derivatives can then be further analyzed by the method of identifying two or more soluble proteins, to determine if fragments of the insoluble proteins are soluble.

[0067] One of ordinary skill in the art will appreciate that the invention provides an improved method for identifying soluble proteins. The inventive method thus will aid in the discovery of cellular targets against which new disease therapeutics can be developed. In

addition, the invention will enable researches to elucidate the molecular pathways underlying specific diseases. The inventive method also will aid in methods for enhancing the solubility of insoluble proteins. In this respect, the deletion derivatives described herein can be used as solubility enhancement tags by creating fusion proteins comprising a soluble deletion derivative and an insoluble protein.

[0068] The following examples further illustrate the invention but, of course, should not be construed as in any way limiting its scope.

EXAMPLES

GENERAL PROCEDURES

Vector Preparation

[0069] 752 individual *C. elegans* open reading frames (ORFs) were obtained from a larger collection of more than 10,000 *C. elegans* ORFs. The complete set of all predicted protein-encoding open reading frames of *C. elegans* (*C. elegans* ORFeome Version 1.1) has been described (see Reboul et al., *Nat. Genet.*, *34*, 35-41 (2003)).

[0070] Using the Gateway® cloning system, the ORFs were incorporated into Gateway® entry plasmid clones (Invitrogen Corp., Carlsbad, CA) (see, e.g., Hartley et al., *Genome Res.*, *10*, 1788-95 (2000) and U.S. Patents 5,888,732, 6,270,969, 6,277,608, and 6,720,140). The Gateway® entry clones typically contain an attL1 site and an attL2 site flanking the ORF of interest, as well as an antibiotic resistance gene, such as a spectinomycin resistance (Sp^R) gene. The DNA concentrations of Gateway® entry clones in eight 96-well plates (i.e., plates 11084, 11085, 11086, 11087, 11094, 11095, 11096, and 11099) were determined by PicoGreen® fluorescence (Molecular Probes, Inc., Eugene, OR), and used to calculate the molar concentration of each plasmid, which ranged from 0 to 8.73 nM. Each 96-well plate contained 94 Gateway® entry clones encoding a *C. elegans* ORF and two control plasmids.

Sample Pooling

[0071] Samples containing less than 0.15 nM of plasmid were omitted from further analyses, thereby leaving 688 ORFs to be pooled. Plasmids were pooled in bins of two-fold concentration range, beginning with the most concentrated plasmid and further subpooled, resulting in a total of six pools. The sub-pools were then combined volumetrically (i.e., one volume of the most concentrated subpool, two volumes of the next most concentrated subpool, etc.). The final pool was ethanol precipitated and dissolved in TE buffer. The predicted proteins encoded by the ORFs in the pool ranged in size from 6.7 kilodaltons (kDa) to 66 kDa.

[0072] For generation of deletion derivatives, plasmid DNA was generated in one of two ways: for liquid pooling, a complete transformation sample was diluted into 100 mL of LB medium supplemented with necessary antibiotics, and grown overnight at 37 °C. For plate-based pooling, the transformation mixture was centrifuged and dissolved in 100 µl LB and plated on 100 mm LB-Agar plates supplemented with necessary antibiotics, and grown overnight at 37 °C. Colonies from the plates were pooled by scraping the plate into 1 mL of LB, which was used immediately to prepare plasmid, or by dilution into a larger volume for overnight growth prior to preparation of plasmid. Plasmid DNA was prepared using the FastPlasmid Mini kit (Eppendorf) per manufacturer's instructions.

Construction of Transposon Donor Plasmids

[0073] The tetracycline resistance gene from the TET-1 transposon (Epicentre Technologies, Madison, WI) was amplified using oligonucleotide primers containing Tn5 mosaic end (ME) sites and Gateway® *attB1* sites. This amplification product was purified using the QiaQuick PCR purification system (Qiagen, Inc., Valencia, CA) and transposed into pUC19 DNA using the Epicentre EZ::TN system according to the manufacturer's instructions. Positive colonies were selected using 100 µg/mL ampicillin and 12.5 µg/mL tetracycline, and plasmid DNA was prepared by standard methods. pUC DNA containing the transposon was sequence verified throughout to confirm the DNA sequence of the entire transposon. To generate large amounts of transposon DNA free of template plasmid, the pUC-transposon DNA was used as a template in a PCR amplification using ME primers. Amplified DNA was digested with DpnI (New England Biolabs, Beverly, MA) for one hour at 37 °C to remove residual plasmid template DNA, heat inactivated by incubation at 80 °C for 20 minutes, and purified using the QiaQuick PCR purification system.

[0074] To generate Gateway® entry plasmid clones containing double deletions, a transposon containing *attB1* sites is used to generate an amino-terminal deletion, while a transposon containing *attB2* sites is used to generate a carboxy-terminal deletion.

Gene Transfer by Electroporation

[0075] Electroporation of DNA was carried out by mixing 1 µl of a DNA mixture with 20 µl of Electromax DH5α-E competent cells (Invitrogen), incubated on ice for 10 minutes, and electroporated in a 0.1 cm gap cuvette using the BioRad (Herecules, CA) MicroPulser according to manufacturer's settings for *Escherichia coli*. After electroporation, samples were diluted in 1 ml of LB or SOC and grown at 200 rpm at 37 °C for 1 hour.

Transposition Reactions

[0076] For transposition reactions, 0.2 μg target DNA and 2-fold molar equivalent of transposon was incubated with 1 unit (U) transposase (Epicentre Technologies, Madison, WI) in 10 μl total volume in 1x transposition buffer (Epicentre Technologies, Madison, WI) for 5-6 hours at 37 °C. The reaction was stopped by addition of 1 μl of 10x stop solution (Epicentre Technologies, Madison, WI) followed by incubation at 70°C for 10 minutes. The reaction was precipitated with one tenth volume of 3 M sodium acetate and 2.5 volumes of 100% ethanol, centrifuged for 20 minutes, dried, and dissolved in 10 μl TE (10 mM Tris-Cl, pH 7.5, 0.1 mM EDTA). 1 μl of the dissolved DNA was electroporated into Electromax DH5α-E cells (Invitrogen, Inc.). After electroporation, samples were diluted in 1 mL of LB, grown at 200 rpm at 37 °C for 1 hour, and plated or diluted for liquid growth using LB supplemented with 100 μg/ml spectinomycin and 12.5 μg/ml tetracycline. Plasmid DNA was prepared as described above.

Selection for Internal Transposition

[0077] To select transposons which insert into the gene of interest, a Gateway® LR reaction was performed using pooled transposed clones and the plasmid pDest-6 (Invitrogen Corp.), which is a Gateway® "Destination Vector" containing an ampicillin resistance (Ap^R) gene, in a 10 μL reaction per the manufacturer's specifications, but with an increased reaction time of 16-18 hours. After stopping the reaction with proteinase K, the reaction was precipitated with ethanol, and dissolved in 10 μl TE. 1 μL of the reaction mixture was electroporated into Electromax DH5α-E cells. After electroporation, samples were diluted in 1 mL of LB medium, grown at 200 rpm at 37 °C for 1 hour, and plated or diluted for liquid growth using LB medium supplemented with 100 μg/mL ampicillin and 12.5 μg/mL tetracycline. Plasmid DNA was prepared as described above.

Generation of Deletion Derivatives and Negative Selection

[0078] To generate Gateway® entry plasmid clones containing deletions, and to select against clones still containing the transposon, pooled transposed expression clones were transferred to a pDonr201 vector via Gateway® BP recombination in a 10 μL reaction per the manufacturer's specifications, but with an increased reaction time of 16-18 hours. The pDonr201 vector contains the toxic *ccdB* gene flanked by *attP* sites (i.e., *attP1* and *attP2*), as well as a kanamycin resistance gene or a spectinomycin resistance gene. 1 μL of the reaction mixture was electroporated into Electromax DH5α-E cells. After electroporation, samples were diluted in 1 mL of LB medium, grown at 200 rpm at 37 °C for 1 hour, and plated or diluted for liquid growth using LB medium supplemented with 50 μg/mL kanamycin and 2 mM nickel sulfate. Plasmid DNA was prepared as described above.

Cell Culture

[0079] Using the Gateway® LR Clonase enzyme mix (Invitrogen Corp.), the 688 pooled ORFs (Gateway® *attL* entry clones) were subcloned into the plasmid pDest527, in which each ORF is expressed as an N-terminal His6 fusion protein under the control of a T7 promoter, according to manufacturer's instructions, except that the reaction was allowed to proceed overnight at room temperature. Reaction products were transformed into DH5α cells (Invitrogen) in SOC standard medium, and 1% of SOC expression mixture was grown on media containing ampicillin. The remaining 99% of the expression mixture was added to 50 mL CircleGrow® media (QBiogene, Inc., Irvine, CA) containing ampicillin, and grown overnight at 37 °C. Plasmid DNA was then purified, and about 100 ng of pooled expression plasmids were electroporated into *E. coli* Rosetta (DE3) strain (Novagen, Madison, WI), resulting in about 1.4 million transformants. The 1 mL of SOC expression mixture was diluted into 50 mL of CircleGrow® containing 100 μg/mL ampicillin and 15 μg/mL chloramphenicol, and grown overnight at 37 °C. The overnight culture was diluted 1:100 into 1L of CircleGrow® containing 100 μg/mL ampicillin and 15 μg chloramphenicol, grown at 37 °C to an A600 concentration of 0.5, and cooled to 16 °C. Protein expression was induced by adding IPTG to 1 mM. After 16 hours at 16 °C, cells were harvested and frozen at -80 °C.

[0080] Deletion derivative clones were transferred to final expression plasmids using the Gateway® LR reaction according to the manufacturer's instructions. 1 μL of the reaction mixture was electroporated into Electromax DH5α-E cells. After electroporation, samples were diluted in 1 mL of LB, grown at 200 rpm at 37 °C for 1 hour, and plated or diluted for liquid growth using LB supplemented with 100 μg/mL ampicillin. Plasmid DNA was prepared as described above. For expression experiments, plasmid DNA was transformed into *E. coli* Rosetta (DE3) cells (Novagen, Madison, WI) and grown in 100 μg/mL ampicillin and 15 μg/mL chloramphenicol. Cells were grown overnight at 37 °C, diluted 1:100 into CircleGrow® media and grown at 37 °C to an OD600 of 0.5. Expression was induced by addition of 0.5 mM IPTG and growth was continued at 37 °C for 2.5 hours or 16 °C overnight. Cells were harvested after induction and 0.05 OD600 units were spun down and dissolved in 1x SDS-loading buffer prior to SDS-PAGE analysis.

Protein Purification

[0081] Unless otherwise noted, the following methods were performed at 4 °C. *E. coli* cell pastes were resuspended with two volumes of extraction buffer per gram of wet weight to achieve a final concentration of 20 mM sodium phosphate buffer, pH 7.5, 100 mM NaCl, 5 mM $MgCl_2$, 5% glycerol, 45 mM imidazole and Complete protease inhibitor with EDTA (Roche Diagnostics Corp., Indianapolis, IN) (1 tablet per 50 mL of extract). Extracts were

treated with lysozyme (0.5 mg/mL) for 30 minutes and with Benzonase® (Novagen, Madison, WI) (10 U/mL) for an additional 20 minutes. Samples were sonicated to lyse the cells (verified by microscopic examination) and adjusted to 500 mM NaCl with solid NaCl. Following centrifugation (111,000 x g for 30 minutes), samples were filtered (0.45 μm, PES membrane), applied to 1 mL HisTrap columns (GE Healthcare, Piscataway, NJ) (0.6 mL/min), and equilibrated with extraction buffer in 500 mM NaCl and 45 mM imidazole (binding buffer). After washing the columns with binding buffer, proteins were eluted with binding buffer and 500 mM imidazole, collected in 1 mL fractions, and analyzed by SDS-PAGE.

[0082] The protein pools were precipitated as follows: (1) trichloroacetic acid was added to 6% (v/v) final concentration and vortexed, (2) five minute incubation on ice, (3) centrifugation at 16,100 x g for 10 minutes, (4) removal of supernatant, (5) five minute incubation of pellets with ice cold acetone, (6) centrifugation at 16,100 x g for five minutes, (7) removal of supernatant, and (8) drying of pellet (two minutes, 70 °C). Precipitated proteins were dissolved in room-temperature solubilization buffer (8 M urea, 4% CHAPS, 50 mM Tris, pH 8.5) to a concentration of 20 mg/mL and stored in 50 μl aliquots at -80 °C.

2D Gel Electrophoresis

[0083] Purified proteins were separated in the first dimension according to their net charge by isoelectric focusing, and were separated in the second dimension according to size (see Figure 6). The nominal isoelectric point (pI) range of the gel was 4 to 7, however the effective pI range for the proteins was about 4 to 6. All the proteins with pIs greater than about 6 appeared along the right edge of the gel. The nominal size range of the gel was 10 to 100 kDa.

[0084] Two-dimensional gel electrophoresis of 200-1000 μg protein was performed according to the procedure described in O'Farrell, *J. Biol. Chem.*, *250*, 4007-21 (1975) with more recent modifications (see, e.g., Link, ed., *2D Proteome Analysis Protocols*, Humana Press (1999)). In the first dimension, isoelectric focusing was accomplished by immobilized pH gradients (IPG) using the Amersham Biosciences/GE Healthcare IPGphor™ isoelectric focusing system. Affinity purified samples were dissolved in 450 μl of rehydration buffer (8M urea, 2% CHAPS, 7 mg DTT, and a trace of bromophenol blue). The rehydration buffer-protein mixture was placed in a 24 cm ceramic strip holder and a 24 cm IPG strip, pH 4-7, was glided, gel side down, into the strip holder. Mineral oil was placed on top of the gel to minimize evaporation and covered with the strip holder plastic cover. The ceramic strip holder was placed in the IPGphor unit for isoelectric focusing under the following conditions: 30 volts for 12 hours, 500 volts for 1 hour, 1000 volts for 1 hour and 8000 volts for 8 hours.

[0085] Prior to the separation in the second dimension, the IPG strip was equilibrated with an SDS buffer system. The equilibration solution contained 50 mM Tris-HCl, pH 8.8, 6 M urea, 30% glycerol, 2% SDS, and a trace of bromophenol blue. Prior to use, 100 mg of DTT was added in 10 mL equilibration buffer. The IPG strips were placed in individual tubes containing the buffer. The tubes were then placed on a rocker and equilibrated for 12 minutes. A second equilibration was performed with 250 mg iodoacetamide solution (instead of DTT) and incubated for another 12 minutes.

[0086] The equilibrated IPG strip was then inserted into a cassette containing a pre-cast Ettan DALT II 12.5% polyacrylamide gel, and contact was made with the 2-D gel. Melted agarose was added to cover the IPG strip. The 2DGE chamber was filled with anode buffer (0.5 M diethanolamine, 0.5 M acetic acid). Cathode buffer (0.1% SDS, 0.192 M glycine, 0.025 M Tris) was added to the top chamber. The running conditions were set in the power supply (phase I, 5 Watts (W) per gel/15min; phase 2, 150W/gel), and electrophoresis continued until the bromophenol blue dye front reached the bottom of the gel (approximately 4-5 hours).

[0087] Once the dye front reached the end of the gel, the cassettes were removed and the gels were placed in Coomassie Brilliant Blue staining solution (25% isopropanol, 10% acetic acid, 0.05% R250 Brilliant Blue) overnight. Gels were then placed in destain solution (30% methanol, 10% acetic acid). All staining/destaining procedures were carried out in glass trays placed on a slowly oscillating rocker table.

[0088] The spots on the 2D gel were numbered 1 to 170, and small pieces were retrieved from the center of each spot and digested with trypsin. The tryptic peptides were separated and sequenced by mass spectrometry. The sequences of the tryptic peptides were compared to predicted sequences of *E. coli* and *C. elegans* proteins in the SwissProt database.

Mass Spectrometry

[0089] Coommassie blue stained protein gel spots were digested as described by Wilm et al., *Nature, 379*, 466-469 (1996). Samples were desalted with C18 Zip Tips (Millipore, Bedford, MA) according to the manufacturer's protocols prior to mass spectrometry (MS) analysis.

[0090] Chromatographic separations were conducted using a 75 μm inner diameter × 360 μm outer diameter × 10 cm long fused silica capillary column (Polymicro Technologies Inc., Phoenix, AZ) with one end flame-pulled to a fine tip. The column was slurry packed with 3 μm, 300 Å pore size C-18 stationary phase (Vydac, Hercules, CA). Microcapillary reverse-phase liquid chromatography (RPLC) was performed using an Agilent 1100 capillary LC system (Agilent Technologies, Palo Alto, CA) coupled online to a linear ion-

trap (LIT) mass spectrometer (LTQ, ThermoElectron, San Jose, CA). Reversed-phase separations were conducted after injecting 5 µl of sample for each analysis. The columns were connected via a stainless steel union to an Agilent 1100 nanoflow LC system (Agilent Technologies, Paolo Alto, CA), which was used to deliver solvents A (0.1% HCOOH in water) and B (0.1% HCOOH in CH_3CN). After sample injection, a 20 minute wash with 98% mobile phase A was applied, and peptides were eluted using a linear gradient of 2% mobile phase B to 42% solvent B over 40 minutes with a constant flow rate of 200 nL/min. The column was washed for 15 minutes with 98% mobile phase B and re-equilibrated with 98% mobile phase A prior to subsequent sample loading.

[0091] The µRPLC column was coupled online to an LIT-MS using the manufacturer's nanoelectrospray source with an applied electrospray potential of 1.5 kV and capillary temperature of 160 °C. The LIT-MS was operated in a data-dependent mode where each full MS scan was followed by five MS/MS scans, in which the five most abundant peptide molecular ions detected from the MS scan were dynamically selected for five subsequent MS/MS scans using a collisional-induced dissociation (CID) energy of 35%.

[0092] The CID spectra were analyzed using SEQUEST operating on a Beowulf 18-node parallel virtual machine cluster computer (ThermoElectron, San Jose, CA) using a combined non-redundant *C. elegans*, *E. coli* proteome database (http://www.expasy.org). Only peptides with conventional tryptic termini (allowing for up to two internal missed cleavages) possessing delta-correlation scores (ΔC_n) > 0.08 and charge state-dependent cross-correlation (X_{corr}) criteria, as follows, were considered as legitimate identifications: >1.9 for +1 charged peptides, >2.2 for +2 charged peptides, and >3.1 for +3 charged peptides.

EXAMPLE 1

[0093] This example demonstrates the use of the inventive method to identify two or more *C. elegans* soluble proteins.

[0094] As described above, 2D gel electrophoresis of the 688 potential *C. elegans* ORFs resulted in 170 spots on the gel. Of the 170 spots, two spots were not processed for mass spectrometry, and three spots were discarded because they contained single peptides from many different proteins, and the spot intensities were weak.

[0095] The remaining 165 spots contained 49 *C. elegans* proteins and 37 *E. coli* proteins. The presence of host proteins in the pool of *C. elegans* proteins purified by virtue of their His6 amino tags was expected, because (a) many proteins have some affinity for immobilized metals (used to purify His6-tagged proteins), and (b) the amount of many *E. coli* proteins applied to the affinity column was much greater than the amount of any single *C. elegans* protein. To further minimize the amount of host cell protein that co-purifies with

the proteins of interest, an additional affinity tag can be used. For example, a Strep2 affinity tag (IBA GmbH, Göttingen, Germany) can be used in conjunction with the His6 affinity tag. The four most intense spots on the gel were *E. coli* proteins and contained three proteins: slyD (proline cis-trans isomerase), dnaK, and groEL. These proteins are stress response proteins presumably induced by overexpression of *C. elegans* proteins in the *E. coli* cells.

[0096] Some *C. elegans* proteins were found in a single spot, while others were found in up to 13 spots. Proteins found in intense spots tended to be found in multiple spots, most of which clustered together. Some spots were pure protein, others contained mixtures of proteins from host *E. coli* cells or *C. elegans*.

[0097] Of the 49 *C. elegans* proteins found by searching predicted *C. elegans* proteins (including those not cloned in the version 1.1 ORFeome), 34 were found in the plates that were combined to make the pooled proteins, and another 5 were found in these plates as "related proteins." Ten proteins were not in the original pool of 688 *C. elegans* ORFs, and of these seven were identified on the basis of a single peptide in a single spot.

[0098] Twelve *C. elegans* proteins were chosen on the basis of both spot intensities and potential biological interest for individual expression (see Table 1). All twelve were found in the subset of the 752 originally screened *C. elegans* ORFs. Each of the twelve ORFs were subcloned into pDest527 and expressed in *E. coli* Rosetta (DE3) cells (700 µl cultures in a 24 well dish). Once the cultures reached an absorbance A600 of 0.5, they were transferred to 17 x 100 mm polypropylene tubes (Falcon 2059), cooled to 16 °C, induced with IPTG (1 mM), and expressed overnight at 16 °C. To examine total expressed protein, 0.1 absorbance units of cells were pelleted, lysed, and applied to a 4 to 20% SDS-PAGE gel (Criterion, Bio-Rad) (see Figure 7). To determine the fraction of soluble and insoluble protein, cells were lysed with detergent, and soluble and insoluble fractions were applied to SDS-PAGE gels (see Figure 8). Soluble fractions were purified through Swell-Gel centrifugal IMAC columns (Pierce Biotechnolgy, Rockford, IL) and analyzed by SDS-PAGE.

[0099] Eight of the twelve ORFs yielded significant soluble protein. The clone of one ORF, ORF #8, was not the predicted size when it was cloned into pDest527, and the cloning was repeated from the well of the plate into pDest527. Two colonies were picked, and the two DNAs from these colonies were transformed into the Rosetta (DE3) *E. coli* expression strain. The second cloning attempt of ORF #8 resulted in a high level of expression (see Figure 9) and a small amount of soluble protein (see Figure 10).

Table 1

Protein Identity	Spot #	Isoelectric Point (pI)	Protein Size (kDa)	His6 Fusion Protein Size (kDa)
Paralyzed arrest at two-fold protein 10 (Troponin C)	29, 43	4.21	18.5	22.7
C. elegans TAP-1 protein (Corresponding sequence C44H4.5) (TAB1-like protein TAP-1)	13, 14, 19, 22, 58, 29, 70, 71, 78, 90, 91, 92, 128	5.05	43.5	47.8
PTL-1A protein (Protein with tau-like repeats protein 1, isoform a). "Related protein"	21, 23, 36, 39	5.01	49.4	53.7
Machado-Joseph disease-like protein	12, 18, 42, 54, 85, 86, 87, 156	5.39	35.9	40.2
Hypothetical protein C17G10.2	31, 48, 49, 96, 138, 144	5.00	44.1	48.4
Skp1p homolog (SKR-12) (Skp1 related (Ubiquitin ligase complex component) protein 12)	15, 47, 139	4.64	18.9	23.2
Hypothetical protein F09G2.9	6, 99, 125,	4.59	44.3	48.6
Bag1 (Human) homolog protein 1 (BAG-family molecular chaperone regulator-1)	2, 41, 166,	5.20	24.0	28.3
Hypothetical protein F53F4.3 in chromosome V	26, 38, 82, 132	4.70	25	29.3
Hypothetical protein D2096.8	33, 34, 107, 108, 110, 111, 112, 113, 115, 121	4.50	35.7	40.0
Probable ATP-dependent RNA helicase p47 homolog	28, 56,	5.55	48.5	52.8
14-3-3-like protein 1 "Related sequence AAA61872" Version 1	59, 60, 130	4.72	28.2	32.5

EXAMPLE 2

[0100] This example demonstrates the use of the inventive method to distinguish soluble proteins from insoluble proteins.

[0101] Twelve positives (i.e., ORFs identified from the most intense spots on the 2D gel) and twelve negatives (i.e., proteins absent from the 2D gel) were isolated from the eight 96-well plates described above. Each ORF was tested individually for expression, solubility, and purification from *E. coli* cells. Specifically, each of the 24 *C. elegans* ORFs was cloned into pDest527, and each of these plasmids was transformed into *E. coli* cells

(Rosetta (DE3) strain, Novagen, Madison, WI). Cultures were grown at 37 °C in 0.7 mL cultures until OD 0.5 was reached. Expression was induced with 1 mM IPTG at 16°C overnight (20 hours). Whole-cell samples were removed, heated with SDS and reducing agent, and applied to an SDS-PAGE gel (gradient 4-20% acrylamide). The remainder of the 700 µl was spun down and lysed using Easy-Lyse™ (Epicentre, Madison, WI). Aliquots of the soluble (supernatant) and insoluble (pellet) fractions were applied to separate gels. The His6-tagged proteins in the remaining soluble fractions were purified by immobilized metal affinity chromatography (IMAC) on Swellgel beads (Pierce Biotechnology, Inc., Rockford, IL) and applied to an SDS-PAGE gel ("IMAC eluants"). The protein concentrations of the 24 IMAC eluants were determined by the Bradford assay.

[0102] The results of this analysis are set forth in Figures 11a-d. Of the twelve positives chosen from the 2D gel, all twelve gave purified proteins visible on a coomassie-stained 1D gel, and seven of these were abundant. Of the twelve negatives, five gave visible purified proteins, and one was abundant. The results of the protein concentration assay (see Bradford et al., *supra*) were consistent with the appearance of the bands on the 1D gel.

[0103] To verify that the above-described experiments could predict successful larger scale behavior, six of the positive ORFs were expressed in *E. coli* individually in one liter cultures. Soluble proteins were released from cells by sonication and ultracentrifugation and purified on preparative IMAC columns. All six ORFs yielded large amounts (i.e., 47 to 374 mg per L) of purified protein.

EXAMPLE 3

[0104] This example demonstrates the use of the inventive method of identifying one or more soluble proteins to determine the subcellular localization of proteins. A vector is engineered to contain a nucleic acid sequence labeled with an affinity tag, and is introduced into a suitable cell as described herein. Following induction of expression of the nucleic acid sequence, subcellular fractions are prepared from the from the host cell. Affinity tagged proteins are purified from the subcellular fractions and are subjected to 2D gel electrophoresis. Proteins in each subcellular fraction are analyzed by mass spectrometry and compared to other 2D gels containing other subcellular fractions, allowing estimation of the distribution of each protein among the various cellular compartments, in parallel, for the entire pool.

EXAMPLE 4

[0105] This example demonstrates using the inventive method of identifying one or more soluble proteins to identify a ligand cell surface binding site. A ligand for a cell surface binding site is immobilized on a solid support. A pool of ORFs is generated and

expressed as part of an Ebna vector in 293E cells as described herein. The pool of cells is passed over the immobilized ligand. Cells bound to the ligand are recovered, and the identity of the ORF mediating binding to the ligand is determined using methods known in the art.

EXAMPLE 5

[0106] This example demonstrates a method of identifying highly expressed but insoluble proteins in accordance with the invention.

[0107] A pellet of insoluble proteins isolated in accordance with the above-described methods is dissolved in 8M Urea, and subjected to 2D gel electrophoresis. The spots expressed at the highest levels represent proteins that are well expressed, but are insoluble. These insoluble proteins can serve as likely targets for solubility enhancement by solubility tags. The insoluble proteins also serve as targets for the generation and analysis of deletion derivatives in accordance with the invention. In this regard, because the full-length protein cannot be expressed in soluble form, fragments of the proteins may serve as useful starting points for structural or functional studies.

[0108] Moreover, comparison of solubility characteristics of the same genes tagged with various solubility enhancers (e.g., maltose-binding protein (MBP), NusA, etc.) may elucidate which solubility tags are the most efficient. The insoluble proteins can be expressed in other heterologous organisms in subsequent experiments using the inventive method. The proteins which are soluble in other organisms may indicate which classes or types of proteins are best expressed in various systems.

EXAMPLE 6

[0109] This example demonstrates the use of the inventive method to produce deletion derivatives of the human folliculin gene.

[0110] Human folliculin is expressed poorly in *E. coli* and has very low solubility. A Gateway® entry clone containing the folliculin gene was used as a target for transposition using the methods described above. The transposon used in this experiment contained only a single *attB1* site, and therefore had inherent directionality. Transposition and selection on plates resulted in 800-2000 tetracycline resistant colonies representing clones which contained a transposon insertion somewhere in the entry clone in one orientation or the other. Clones were sequenced using transposon specific primers to identify the site and orientation of insertion. Sixteen clones were chosen and pooled together. Of these clones, eleven were identified as transposition events into the folliculin gene. Three clones were identified as in-frame correct orientation insertions, five clones were out-of-frame correct orientation insertions, and three clones were in the reverse orientation.

[0111] The pool of 16 clones was recombined by Gateway® LR recombination into a Gateway® destination plasmid conferring ampicillin resistance and selected on medium containing ampicillin and tetracycline to eliminate transposition events outside of the folliculin gene. Pooled DNA from these clones was subsequently transferred to a pDonr223 plasmid by Gateway® BP recombination in order to subclone portions (i.e., fragments) of the folliculin gene from plasmids containing the transposon in the desired orientation. The final pool of deleted Entry clones were plated on spectinomycin and nickel sulfate, and colonies were picked for sequence analysis to verify the presence of the deletions. This pool was then transferred to a pDest-586 plasmid by Gateway® LR recombination to produce a final pool of expression clones with amino terminal His6-MBP fusions. At the same time, the three in-frame individual clones from the pool were transferred to pDest-586 individually. The individual clones and the pooled clones were transformed into Rosetta (DE3) *E. coli* cells (Epicentre Technologies, Madison, WI), and expressed.

[0112] Expression of the individual clones produced proteins of the expected molecular weight for the His6-MBP fusions to the deletion fragments. Two of the three individual in-frame fragments expressed significant levels of protein, while the third produced a much lower level of protein and induced significant toxicity in the cells, as seen by a dramatic decrease in growth rate post-induction. The pooled sample showed induction of at least six protein bands, including bands which corresponded to the three individual proteins, and others which were the expected size for several of the out-of-frame deletions.

EXAMPLE 7

[0113] This example demonstrates the use of the inventive method to construct amino terminal deletions of the human folliculin gene.

[0114] A Gateway® entry clone containing the folliculin gene was used as a target for transposition using the methods described above. The transposon used in this experiment contained *attB1* sites at both ends to permit productive recovery of deletions from all transposition events. Transposition and selection on plates resulted in more than 2000 Tet-resistant colonies representing clones which contained a transposon insertion somewhere in the entry clone in one orientation or the other. A pool of these clones was used to transfer the transposed genes into an expression vector to select for transposition events in the gene of interest. The clones were then pooled and transferred back into entry clones by recombination, and plated on selective media to select for deletions. 96 clones were selected and DNA was prepared and sequenced using transposon specific primers to identify the site and orientation of insertion. Approximately one-third of the clones contained deletions in the proper reading frame, as expected. A pool of these clones was transferred to a pDest-586 plasmid by Gateway® LR recombination to produce a final pool

of expression clones with amino terminal His6-MBP (maltose binding protein) fusions. At the same time, 12 in-frame individual clones from the pool were transferred to pDest-586. The individual clones and the pooled clones were transformed into Rosetta (DE3) cells, and expressed.

[0115] Expression of the individual clones produced proteins of the expected molecular weight for the His6-MBP fusions to the deletion fragments. The pooled sample showed induction of at least six protein bands, including bands which corresponded to known individual proteins, and others which were the expected size for several of the out-of-frame deletions. The solubility behavior of these deletion clones correlated well with data previously obtained from studies of specific deletions constructed by PCR.

EXAMPLE 8

[0116] This example demonstrates a method of determining the solubility of protein deletion derivatives.

[0117] A pool of deletion derivatives produced using the methods in Example 6 are screened by transferring the pool to specially designed expression clones containing carboxy terminal fluorescent protein fusions. The fluorescence of the partner protein is linked to the solubility of the deletion fragment (see, e.g., Waldo et al., *Nat. Biotechnol.*, *17*, 691-695 (1999)). Therefore, only clones which are both in-frame and soluble produce fluorescent colonies, which can be easily screened. Soluble deletion derivatives also can be screened by transferring a pool of deletion derivatives to an expression vector which contains a carboxy terminal fusion with an antibiotic resistance gene (e.g., chloramphenicol acetyl transferase (CAT)). In-frame, soluble fusions produce a protein that provides resistance to chloramphenicol, and data shows that levels of resistance correlate with levels of solubility (see, e.g., Maxwell et al., *Protein Science*, *8*, 1908-1911 (1999)). This provides a direct readout of solubility of the deletion derivatives. In addition, a pool of deletion derivatives can be screened using the method described in Example 3 by transferring the pool to a suitable expression vector, followed by pooled growth and purification. Samples can then be separated on 2D gels to identify soluble fragments.

[0118] A pool of deletion derivatives also can be used as solubility enhancement tags. For example, large numbers of deletion derivatives of proteins of interest can be screened to identify proteins which have the best chance of enhancing the solubility of their partners. To this end, new destination vectors can be developed which produce carboxyterminal fusions to strongly insoluble proteins (e.g., ketosteroid isomerase (KSI)). The pool of deletion fragments fused to the insoluble protein can then be screened using the methods described in Examples 1-3 to identify fragments which provide the highest level of solubility enhancement.

[0119] While the inventive methods preferably are practiced *in vivo*, the inventive methods can also be practiced *in vitro* with only minor modifications that are well within the skill in the art.

[0120] All references, including publications, patent applications, and patents, cited herein are hereby incorporated by reference to the same extent as if each reference were individually and specifically indicated to be incorporated by reference and were set forth in its entirety herein.

[0121] The use of the terms "a" and "an" and "the" and similar referents in the context of describing the invention (especially in the context of the following claims) are to be construed to cover both the singular and the plural, unless otherwise indicated herein or clearly contradicted by context. The terms "comprising," "having," "including," and "containing" are to be construed as open-ended terms (i.e., meaning "including, but not limited to,") unless otherwise noted. Recitation of ranges of values herein are merely intended to serve as a shorthand method of referring individually to each separate value falling within the range, unless otherwise indicated herein, and each separate value is incorporated into the specification as if it were individually recited herein. All methods described herein can be performed in any suitable order unless otherwise indicated herein or otherwise clearly contradicted by context. The use of any and all examples, or exemplary language (e.g., "such as") provided herein, is intended merely to better illuminate the invention and does not pose a limitation on the scope of the invention unless otherwise claimed. No language in the specification should be construed as indicating any non-claimed element as essential to the practice of the invention.

[0122] Preferred embodiments of this invention are described herein, including the best mode known to the inventors for carrying out the invention. Variations of those preferred embodiments may become apparent to those of ordinary skill in the art upon reading the foregoing description. The inventors expect skilled artisans to employ such variations as appropriate, and the inventors intend for the invention to be practiced otherwise than as specifically described herein. Accordingly, this invention includes all modifications and equivalents of the subject matter recited in the claims appended hereto as permitted by applicable law. Moreover, any combination of the above-described elements in all possible variations thereof is encompassed by the invention unless otherwise indicated herein or otherwise clearly contradicted by context.

WHAT IS CLAIMED IS:

1. A method for identifying each of two or more soluble proteins, which method comprises:

(a) preparing a mixture of two or more vectors each comprising a nucleic acid sequence encoding a soluble protein operatively linked to a promoter,

(b) transferring each of the two or more vectors into each of one or more cells,

(c) expressing the nucleic acid sequence in the one or more cells, wherein the two or more soluble proteins are produced,

(d) purifying the two or more soluble proteins from the one or more cells,

(e) separating the two or more soluble proteins,

(f) isolating each of the two or more soluble proteins, and

(g) subjecting the two or more soluble proteins to mass spectrometry, whereupon (i) each of the two or more soluble proteins is identified, and (ii) the amount of each of the two or more soluble proteins produced in the one or more cells is determined.

2. The method of claim 1, wherein each of the two or more soluble proteins is derived from the same organism.

3. The method of claim 1, wherein each of the two or more soluble proteins is derived from different organisms.

4. The method of any of claims 1-3, wherein the two or more soluble proteins are separated by electrophoresis.

5. A method for producing soluble deletion derivatives of a protein, which method comprises:

(a) preparing a vector comprising a nucleic acid sequence encoding the protein, wherein the nucleic acid sequence is flanked by a first and a second site-specific recombination site, and wherein the first and second site-specific recombination sites do not recombine with each other,

(b) incubating the vector of (a) in the presence of one or more transposons and a transposase protein under conditions sufficient to cause insertion of the one or more transposons into the vector, wherein each of the one or more transposons comprises a third and a fourth site-specific recombination site,

(c) transferring the nucleic acid sequence into a second vector, wherein the second vector comprises a fifth and a sixth site-specific recombination site, and propagating the second vector in a bacterium,

(d) isolating one or more second vectors of (c) comprising the one or more transposons inserted into the nucleic acid sequence,

(e) combining the second vector of (d) with a third vector to produce a mixture, wherein the third vector comprises a seventh and an eighth site-specific recombination site, and

(f) incubating the mixture of (e) in the presence of at least one recombinase protein under conditions sufficient to cause recombination of (i) the seventh and eighth site-specific recombination sites of the third vector with (ii) the fourth site-specific recombination site of one transposon and the sixth site-specific recombination site of the second vector, wherein one or more nucleotides of the nucleic acid sequence are deleted, and whereupon one or more soluble deletion derivatives of the protein are produced.

6. The method of claim 5, wherein the protein is a mammalian protein.

7. The method of claim 6, wherein the protein is a human protein.

8. The method of any of claims 5-7, wherein each of the one or more transposons is a Tn5 transposon.

9. A method for identifying one or more protein portions in a sample, which method comprises:

(a) preparing two or more soluble deletion derivatives of one or more proteins according to the method of claim 5,

(b) separating the two or more soluble deletion derivatives,

(c) isolating each of the two or more separated soluble deletion derivatives,

(d) subjecting each of the two or more soluble deletion derivatives to mass spectrometry, whereupon the one or more protein portions are identified.

10. The method of claim 9, wherein the nucleic acid sequence encoding each of the two or more deletion derivatives comprises a different deletion of one or more nucleotides of the nucleic acid sequence

11. The method of claim 9, wherein each of the one or more proteins are derived from the same organism.

12. The method of claim 9, wherein each of the one or more proteins are derived from different organisms.

13. The method of any of claims 9-12, wherein the two or more deletion derivatives are separated by electrophoresis.

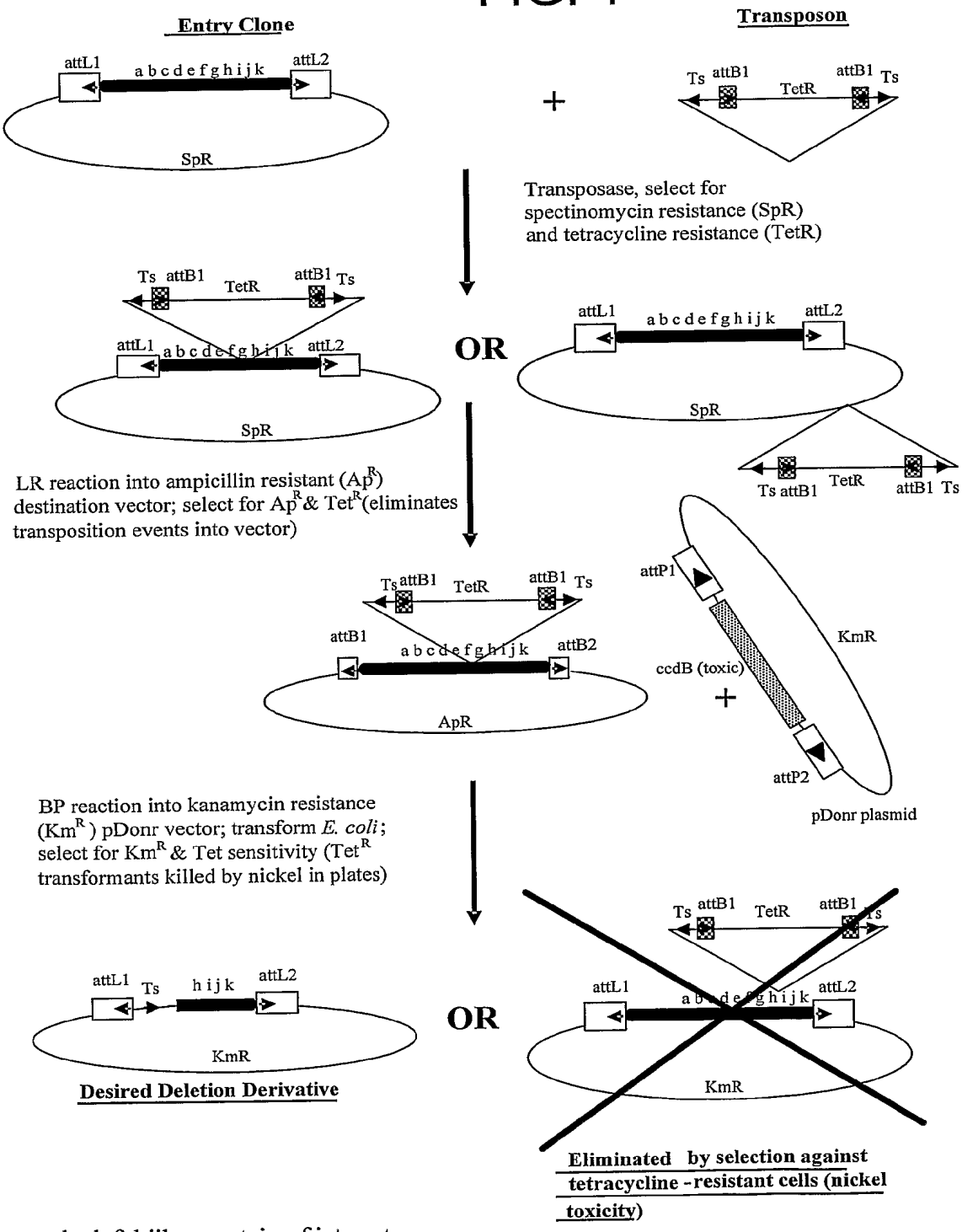
14. A deletion derivative of a protein produced by the method of any of claims 5-13.

15. The method of any of claims 2, 3, 11, or 12, wherein the organism is a mammal.

16. The method of claim 15, wherein the mammal is a human.

17. A nucleic acid molecule comprising a transposon, which transposon comprises a nucleic acid sequence encoding a selectable marker flanked by identical inverted site-specific recombination sequences.

FIG. 1



abcdefghijk = protein of interest
Ts = Tn5 transposase recognition sequence

FIG. 2

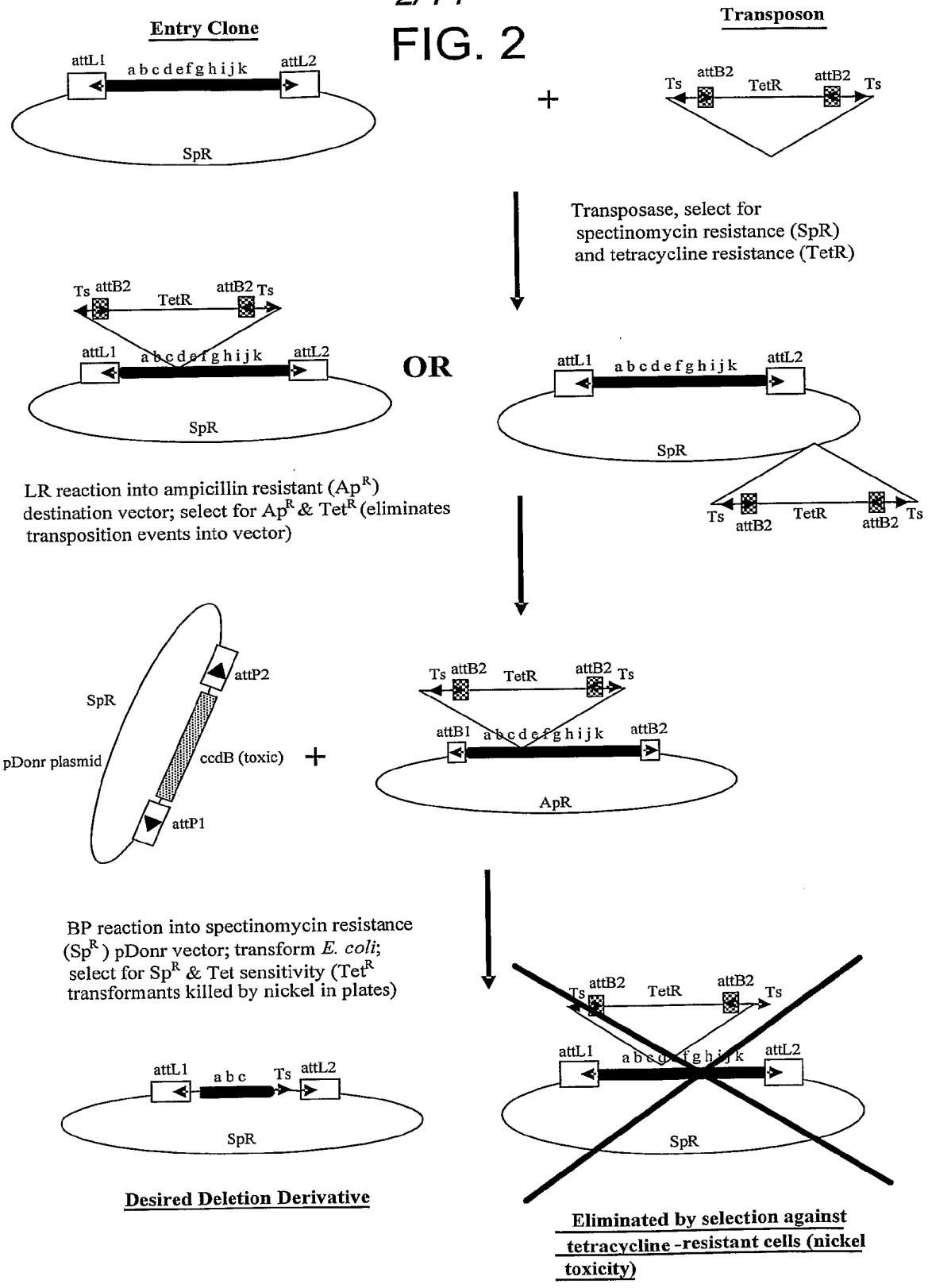

Entry Clone
attL1
a b c d e f g h i j k
attL2
SpR
+
Transposon
Ts
attB2
TetR
attB2
Ts
Transposase, select for spectinomycin resistance (SpR) and tetracycline resistance (TetR)
OR
LR reaction into ampicillin resistant (Ap^R) destination vector; select for Ap^R & Tet^R (eliminates transposition events into vector)
attP2
SpR
pDonr plasmid
ccdB (toxic)
attP1
attB1
ApR
BP reaction into spectinomycin resistance (Sp^R) pDonr vector; transform E. coli; select for Sp^R & Tet sensitivity (Tet^R transformants killed by nickel in plates)
a b c
Desired Deletion Derivative
Eliminated by selection against tetracycline -resistant cells (nickel toxicity)

FIG. 3

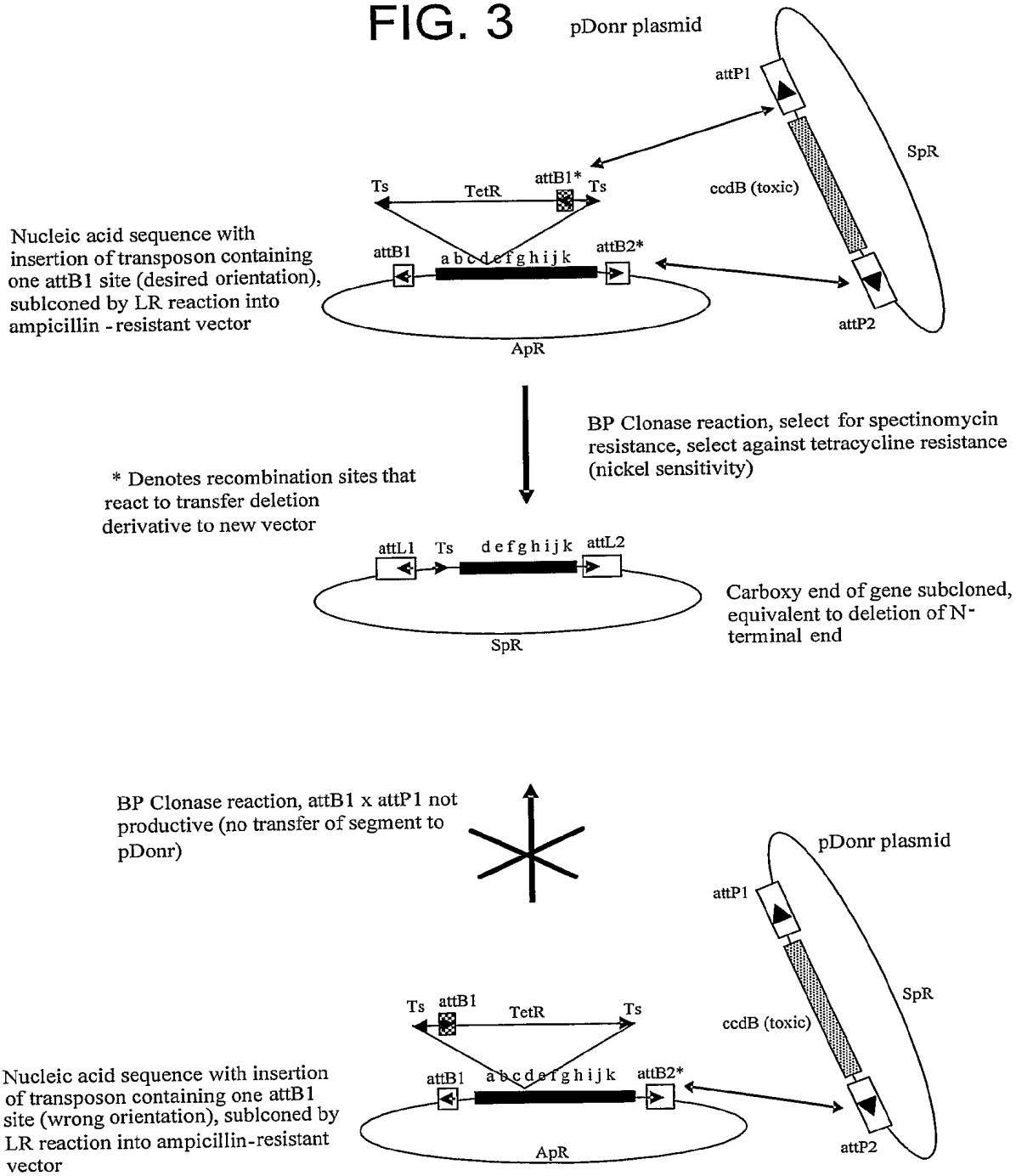

pDonr plasmid
attP1
SpR
Ts
TetR
attB1*
Ts
ccdB (toxic)
Nucleic acid sequence with insertion of transposon containing one attB1 site (desired orientation), subIconed by LR reaction into ampicillin - resistant vector
attB1
a b c d e f g h i j k
attB2*
attP2
ApR
BP Clonase reaction, select for spectinomycin resistance, select against tetracycline resistance (nickel sensitivity)
* Denotes recombination sites that react to transfer deletion derivative to new vector
attL1
Ts
d e f g h i j k
attL2
SpR
Carboxy end of gene subcloned, equivalent to deletion of N-terminal end
BP Clonase reaction, attB1 x attP1 not productive (no transfer of segment to pDonr)
pDonr plasmid
attP1
SpR
Ts
attB1
TetR
Ts
ccdB (toxic)
Nucleic acid sequence with insertion of transposon containing one attB1 site (wrong orientation), subIconed by LR reaction into ampicillin-resistant vector
attB1
a b c d e f g h i j k
attB2*
ApR
attP2

FIG. 4

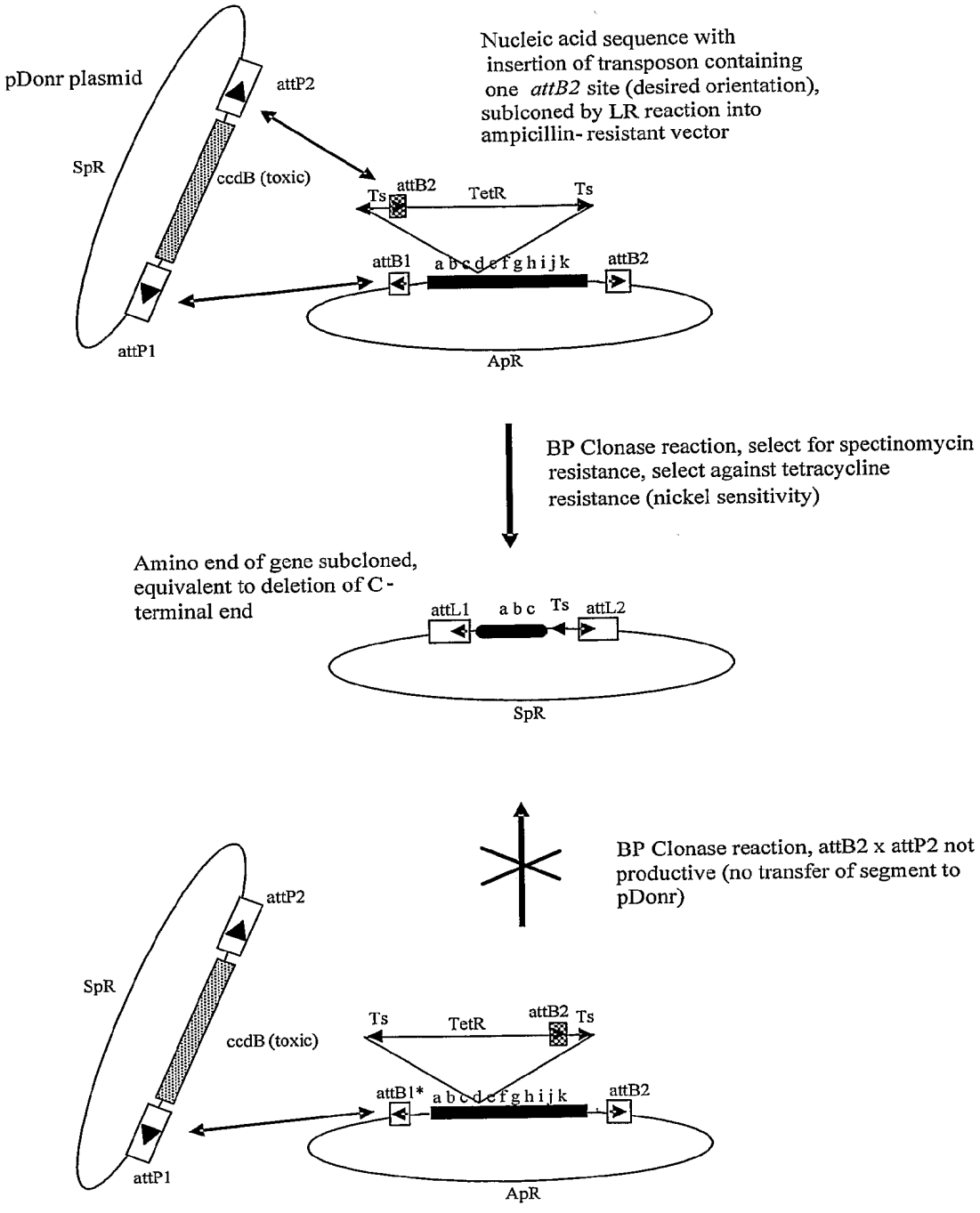
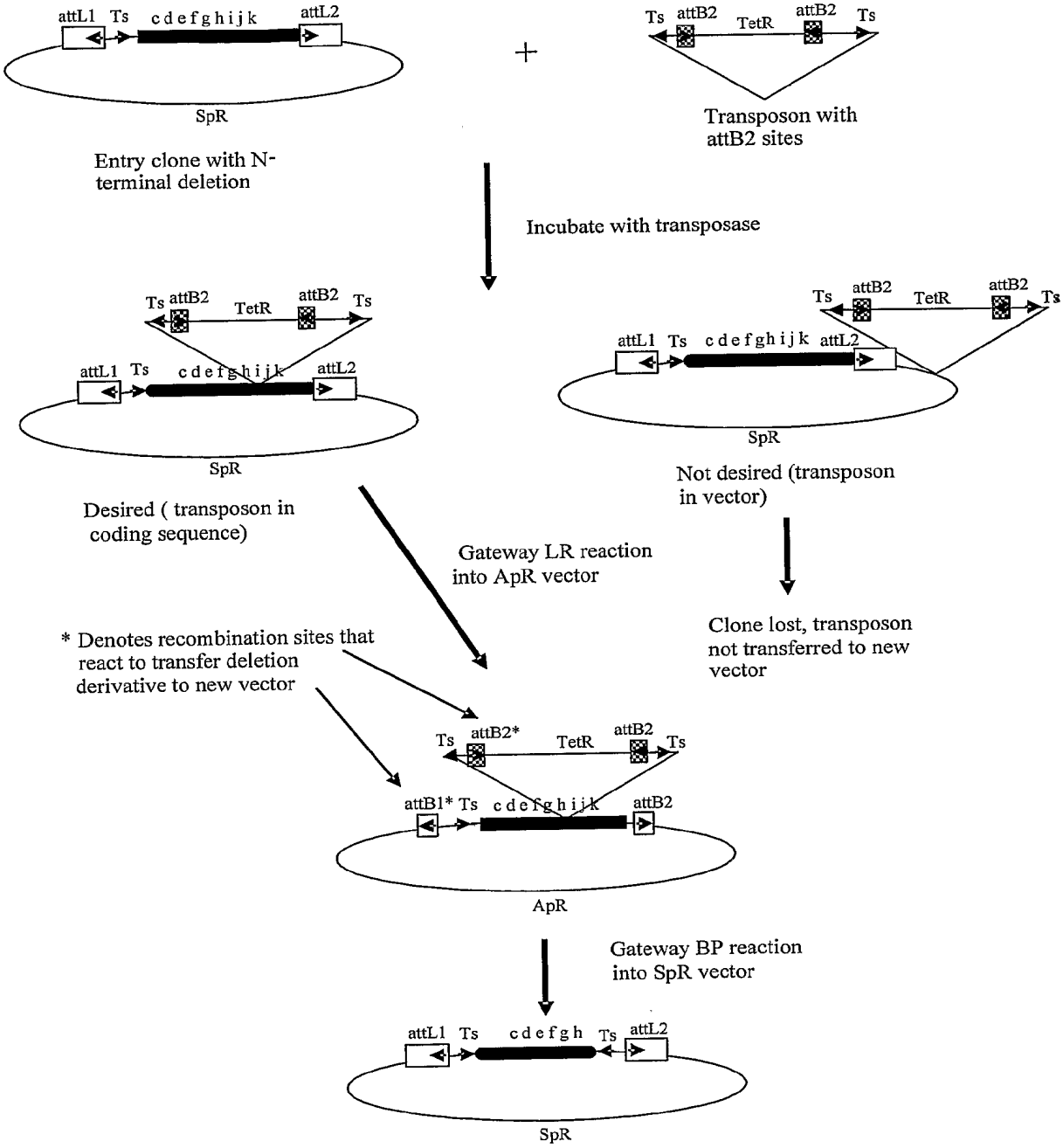

pDonr plasmid
attP2
SpR
ccdB (toxic)
attP1
Nucleic acid sequence with insertion of transposon containing one attB2 site (desired orientation), sublconed by LR reaction into ampicillin-resistant vector
Ts
attB2
TetR
Ts
attB1
a b c d e f g h i j k
attB2
ApR
BP Clonase reaction, select for spectinomycin resistance, select against tetracycline resistance (nickel sensitivity)
Amino end of gene subcloned, equivalent to deletion of C-terminal end
attL1
a b c
Ts
attL2
SpR
BP Clonase reaction, attB2 x attP2 not productive (no transfer of segment to pDonr)
attP2
SpR
ccdB (toxic)
attP1
Ts
TetR
attB2
Ts
attB1*
a b c d e f g h i j k
attB2
ApR
Nucleic acid sequence with insertion of transposon containing one attB2 site (wrong orientation), sublconed by LR reaction into ampicillin-resistant vector

FIG. 5



Double deletion derivative: N-terminal deletion via attB1 transposon, C-terminal deletion via attB2 transposon.

FIG. 6

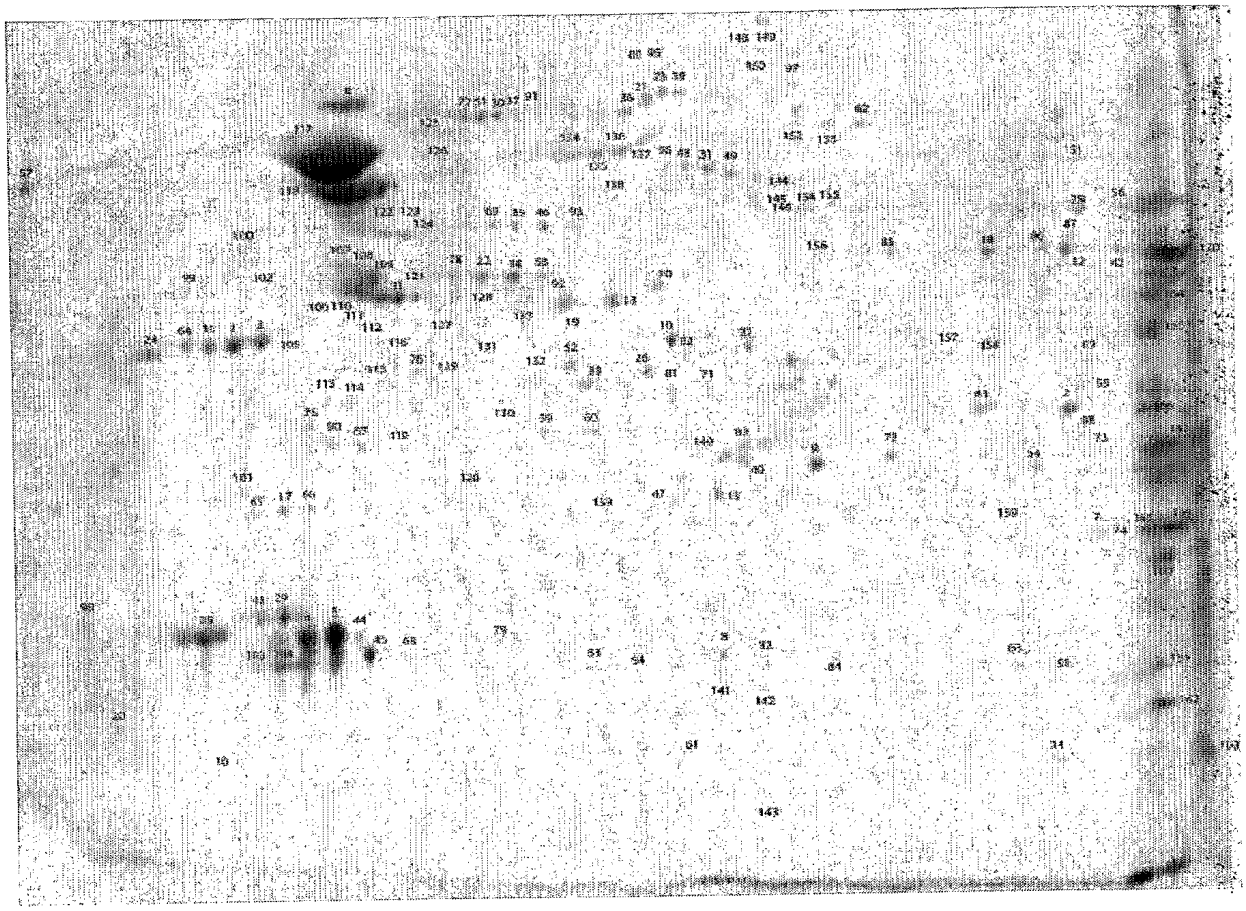


FIG. 7

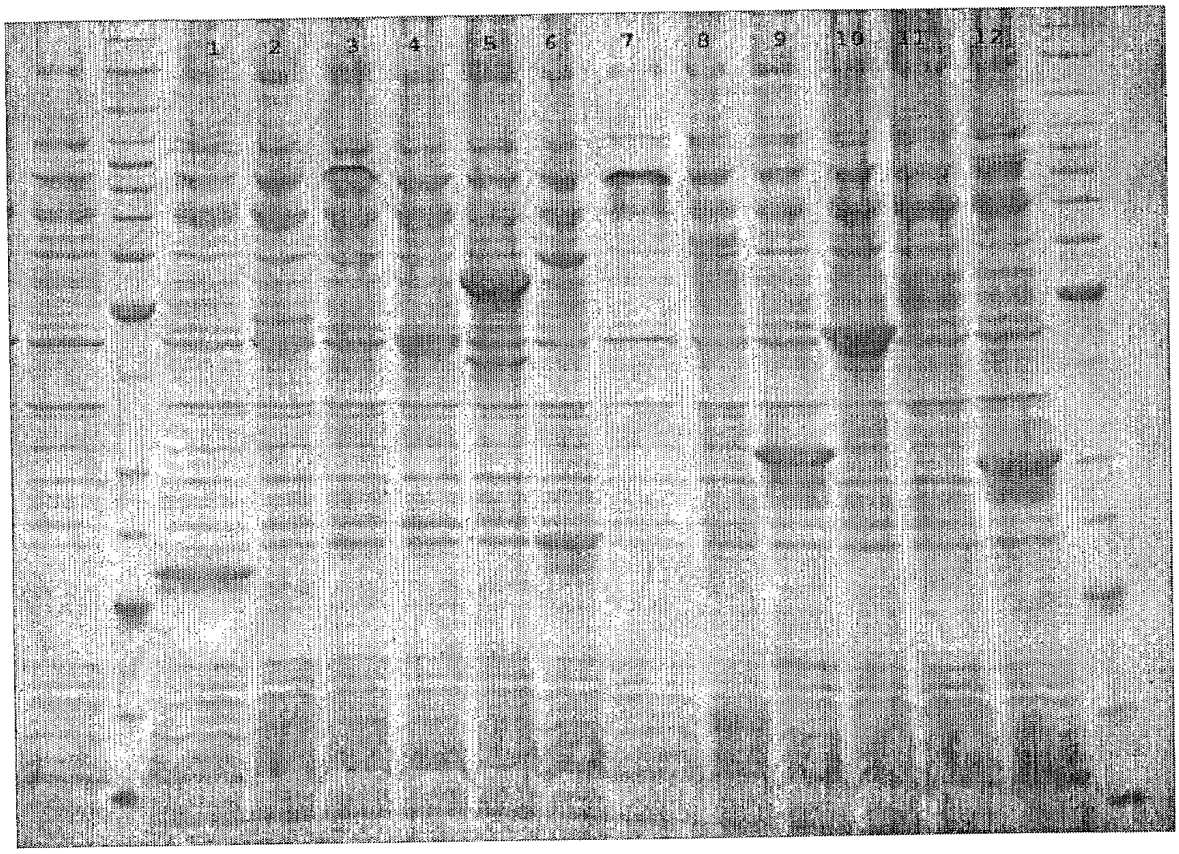


FIG. 8

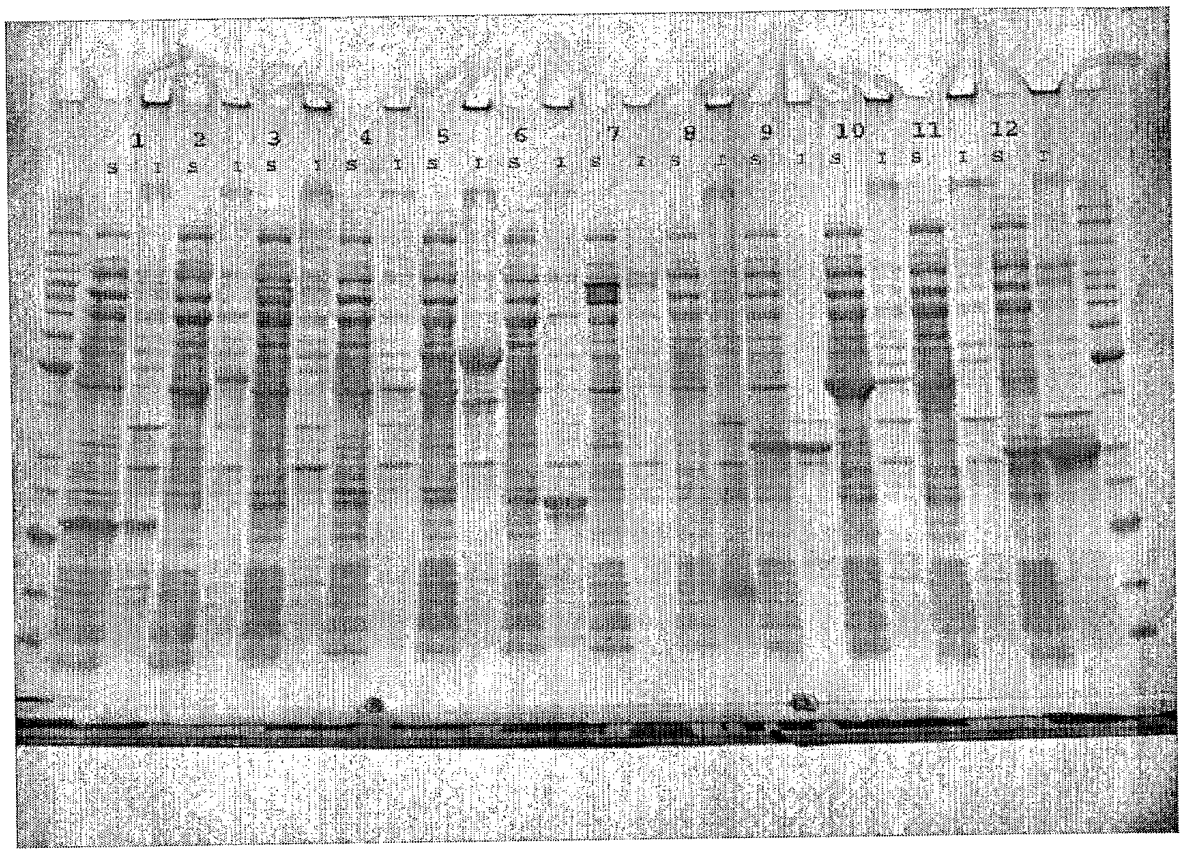


FIG. 9

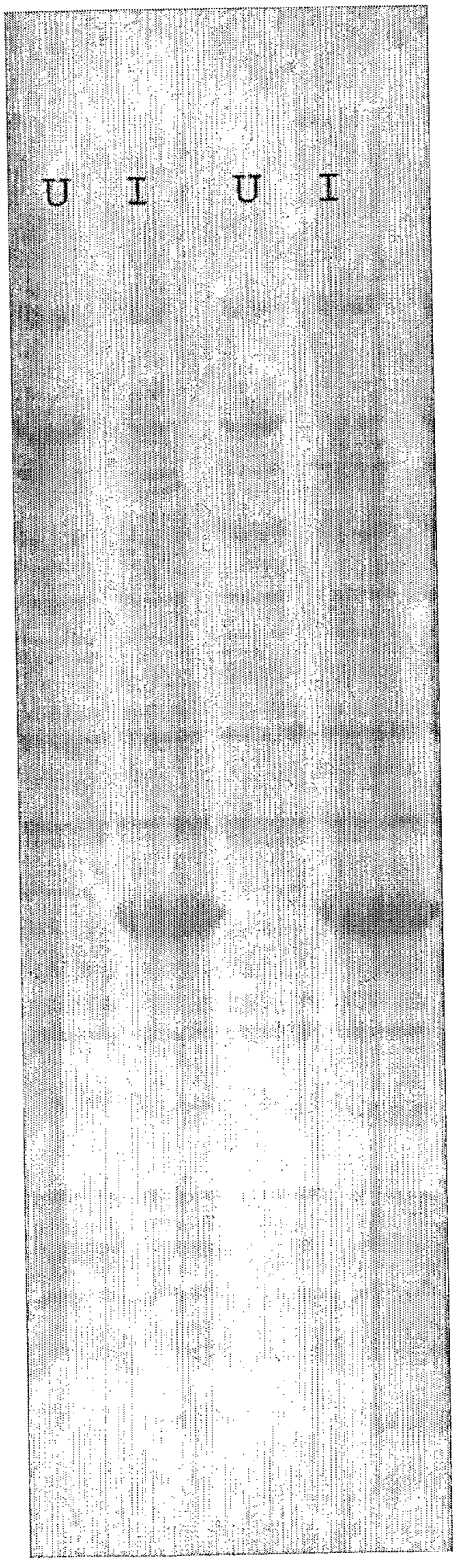


FIG. 10

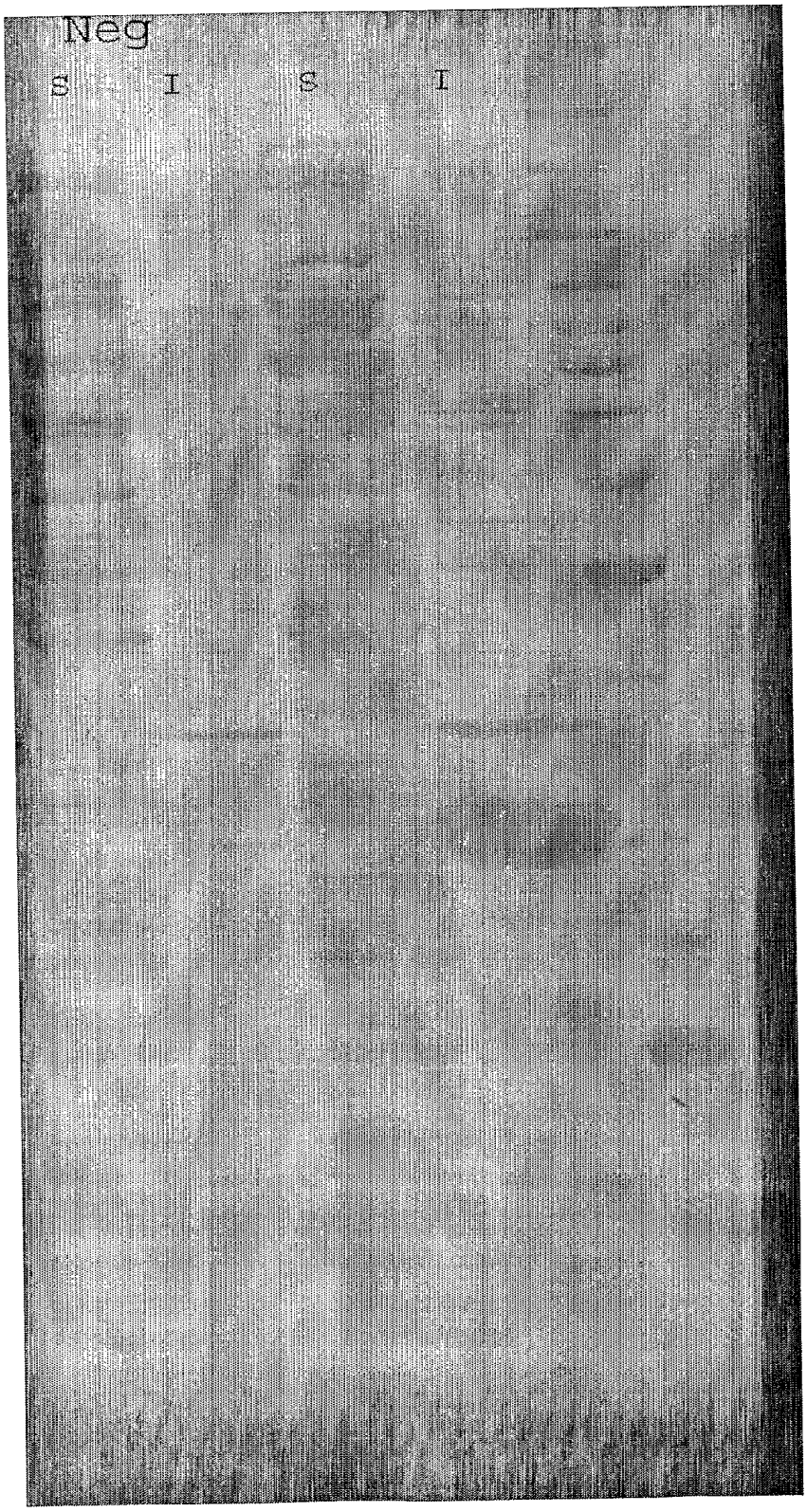


FIG. 11A

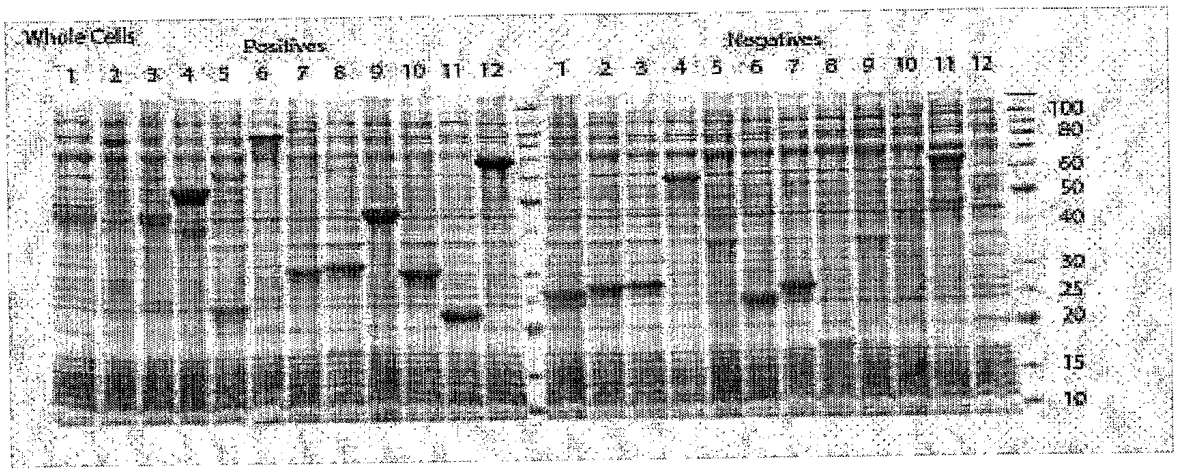

Whole Cells
Positives
Negatives
1 2 3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 7 8 9 10 11 12
100
80
60
50
40
30
25
20
15
10

FIG. 11B

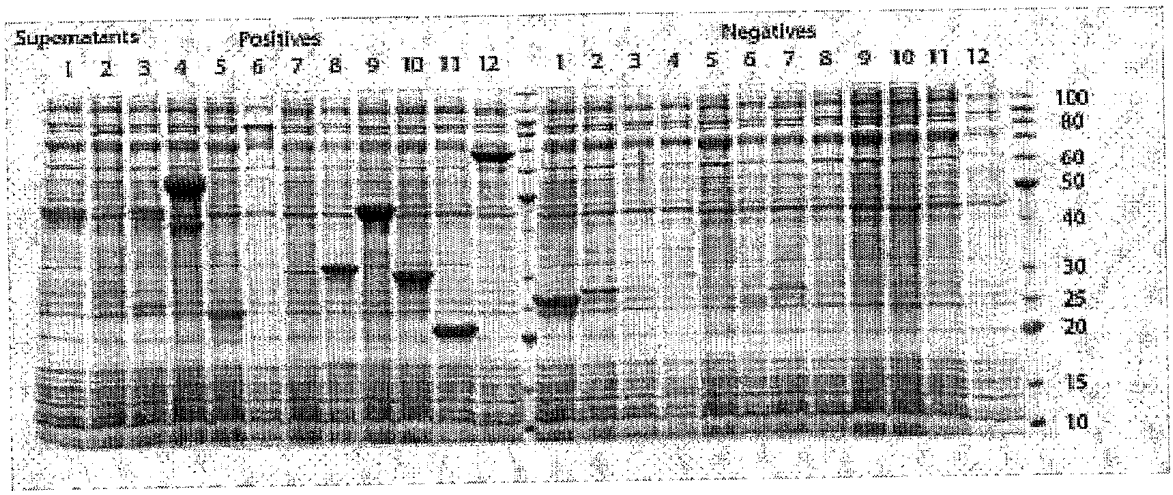

Supernatants
Positives
Negatives
1 2 3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 7 8 9 10 11 12
100
80
60
50
40
30
25
20
15
10

FIG. 11C

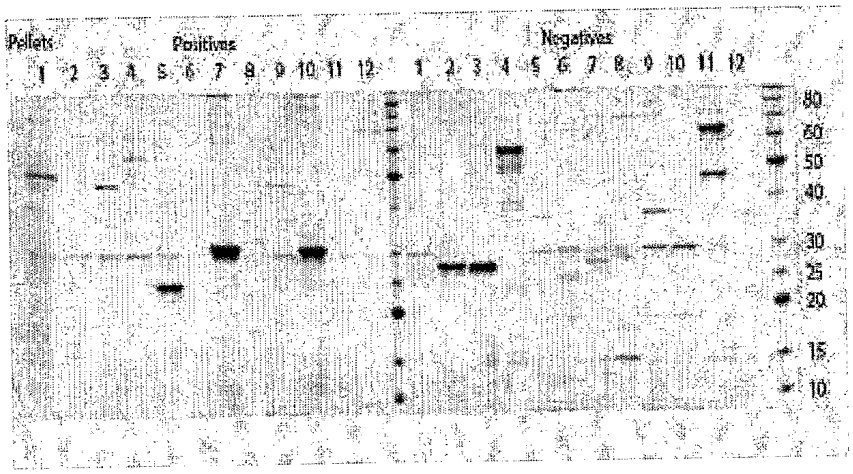
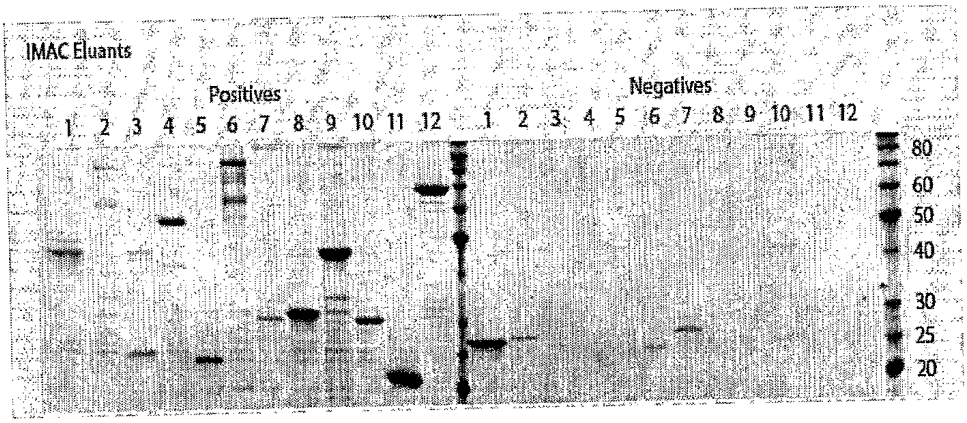

Pellets
Positives
Negatives
1 2 3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 7 8 9 10 11 12
80
60
50
40
30
25
20
15
10

FIG. 11D


IMAC Eluants
Positives
Negatives
1 2 3 4 5 6 7 8 9 10 11 12
1 2 3 4 5 6 7 8 9 10 11 12
80
60
50
40
30
25
20